**Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache**

**Corpus Linguistics and Interdisciplinary Perspectives on Language**

Bd./Vol. 4

**Noah Bubenhofer**
**Marek Konopka**
**Roman Schneider**

# Präliminarien einer Korpusgrammatik

Unter Mitwirkung von
Caren Brinckmann, Katrin Hein und Bruno Strecker

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.



Internet: www.narr.de
E-Mail: info@narr.de

Redaktion: Elke Donalies, Mannheim
Layout: Andreas Scholz, Essen; www.a-shots.de
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6701-7

# Inhalt

# Einleitung

Dieser Band soll die ersten Schritte auf dem Weg zu einer neuen, korpuslinguistisch fundierten Grammatik des Standarddeutschen dokumentieren, die am Institut für Deutsche Sprache in Mannheim (IDS) geplant ist. In dieser „Korpusgrammatik" soll der Idee, dass die Standardsprache nichts Starres ist, sondern in Abhängigkeit von einer Reihe von Faktoren variiert, eine besondere Rolle zukommen, denn die korpuslinguistische Methodik gibt uns heute die Chance, die bisher oft übersehene Vielfalt und Variabilität des standardsprachlichen Sprachgebrauchs detaillierter und zuverlässiger denn je zu analysieren.

Dabei beschränken sich die Einsatzgebiete quantitativer Korpusforschung selbstverständlich nicht nur auf den grammatischen Bereich, sondern umfassen Fragestellungen aus vielen sprachwissenschaftlichen Teildisziplinen wie Lexikologie, Pragmatik oder Semantik. Vor diesem Hintergrund sind auch die weltweit erkennbaren Anstrengungen zum Aufbau (zunehmend mehrfach) annotierter umfangreicher Sprachkorpora zu sehen, die sich beispielsweise in folgenden Projekten manifestieren:

- Bulgarisch: *Bulgarian National Corpus* (http://ibl.bas.bg/en/BGNC_en.htm), mit ca. 1,2 Milliarden laufenden Wortformen;
- Deutsch: *Deutsches Referenzkorpus* (*DeReKo*; http://www.ids-mannheim.de/kl/projekte/korpora/), ca. 6 Milliarden Wortformen;
- Englisch: *Corpus of Global Web-based English* (*GloWbE*; http://corpus2.byu.edu/glowbe/), ca. 1,9 Milliarden Wortformen;
- Französisch: *Korpus I-FR* (http://Corpus.Leeds.ac.uk/internet.html), ca. 200 Millionen Wortformen;
- Italienisch: *PAISÀ-Korpus* (http://www.corpusitaliano.it), ca. 250 Millionen Wortformen;
- Polnisch: *National Corpus of Polish* (NKJP; http://nkjp.pl), ca. 1,8 Milliarden Wortformen;
- Russisch : *Russisches Nationalkorpus* (http://ruscorpora.ru), ca. 300 Millionen Wortformen;
- Schwedisch: *Språkbanken* (http://spraakbanken.gu.se), ca. 1,3 Milliarden Wortformen;
- Slowenisch: *GigaFIDA* (http://www.gigafida.net), ca. 1,2 Milliarden Wortformen;

- Spanisch: *Corpus del Español Actual* (*CEA*; http://sfncorpora.uab.es/CQPweb/cea/), ca. 540 Millionen Wortformen;
- Tschechisch: *Cesky Národní Korpus* (*CNK*; http://ucnk.ff.cuni.cz/english/index.php), ca. 1,3 Milliarden Wortformen

Diese – zugegebenermaßen unvollständige[1] – Übersicht illustriert die Hoffnungen, die sich in der internationalen linguistischen Forschung mit dem Einsatz quantitativer Methoden verbinden. Und abgesehen von grundlegenden Voraussetzungen und Kriterien, die für eine angemessene Benutzung derartiger Sprachkorpora stets zu beachten sind – Ausgewogenheit, Repräsentativität und Aktualität der Textquellen, Qualität der Annotationen, Leistungsfähigkeit der Recherchesysteme, um nur einige zu nennen – findet sich in dieser Liste eine wichtige Ermutigung für unser Vorhaben: Korpusbasierte Untersuchungen zur Grammatik des Deutschen können sich auf eine vergleichsweise breite Datenbasis stützen, die darüber hinaus auch über vielfältige Metainformationen (Entstehungszeiten, Textarten, Sachgebiete) erschließbar ist.

Doch ist eine korpuslinguistische Herangehensweise an Grammatik einfach nur Erbsenzählerei? Natürlich spielt das Auszählen von Phänomenen im Korpus eine zentrale Rolle, doch die Korpuslinguistik kann auf eine reiche Tradition der Statistik und der quantitativen Linguistik zurückgreifen, die einen Methodenapparat hervorbrachte, der weit mehr ist als Zählen.

Die Statistik unterscheidet deskriptive, induktive und explorative Verfahren. In der deskriptiven Statistik werden tatsächlich „Erbsen" gezählt, allerdings auf hohem methodischen Niveau. Denn neben den Frequenzen eines Phänomens in einem Korpus und den Mittelwerten über mehrere Korpora hinweg gibt es eine Reihe von Methoden, um die Verteilung und Streuung der Phänomene in den Daten zusammenfassen zu können. Es ist beispielsweise einleuchtend, dass die pure Auftretensfrequenz eines Phänomens in einem Korpus keine hohe Aussagekraft hat. Gerade bei eher niedrigen Frequenzen stellt sich die Frage, ob sich das Phänomen auf ein paar wenige Texte im Korpus beschränkt oder aber gleichmäßig im Korpus verteilt ist. Ein Verteilungsmaß

[1] Unsere Auflistung berücksichtigt synchron ausgerichtete wissenschaftliche Korpora europäischer Sprachen und nennt jeweils ein exemplarisches Beispiel. Weitere Informationen bieten z.B. die Kataloge der *European Language Resources Association ELRA* (http://catalog.elra.info) oder des *Linguistic Data Consortium LDC* (http://www.ldc.upenn.edu/Catalog/), die Korpusliste der *CLARIN-D*-Zentren (http://de.clarin.eu/de/sprachressourcen/corpora), die Korpusliste des Tübinger SFB 441 (http://u-002-ssfbv001.uni-tuebingen.de/sfb441/c1/corpora.html) sowie das Ressourcenverzeichnis des Computerlinguistik-Portals (http://www.computerlinguistik.org).

wie Gries' *Deviation of Proportions* (DP)[2] drückt aus, wie gleichmäßig ein Phänomen verteilt ist. In Kombination mit den Frequenzwerten ergibt sich dadurch ein besseres Bild der beobachteten Daten.

Die induktive Statistik versucht noch einen Schritt weiter zu gehen: Sie versucht mit geeigneten statistischen Tests zu überprüfen, ob beispielsweise Frequenzunterschiede eines Phänomens zwischen verschiedenen Teilkorpora statistisch signifikant sind, bzw. ob mit einer genügend großen Sicherheit die sogenannte Nullhypothese, dass der Unterschied rein zufällig entstanden ist, abgelehnt werden kann. Dabei werden die unterschiedlichen Größen der Teilkorpora mit berücksichtigt, so dass es einfacher wird, beobachtete Frequenzunterschiede als linguistisch erklärungsbedürftig oder nicht weiter bedeutsam (da im Rahmen zu erwartender Schwankungen) zu klassifizieren. Diese Tests funktionieren zudem sowohl für sehr frequente als auch niedrigfrequente Phänomene.

Schließlich sind auch Methoden der explorativen Statistik für korpusgrammatische Fragestellungen interessant. Mit deren Hilfe wird versucht, Zusammenhänge zwischen Daten zu erkennen und daraus Hypothesen abzuleiten. So wird beispielsweise explorativ vorgegangen, wenn über ein Verfahren des maschinellen Lernens die Bedingungen ergründet werden sollen, die bei nominalen Komposita zu den spezifischen Fugen führen. Welche Kombinationen von Faktoren (Suffix oder Auslaut des Erstglieds, Flexionsklasse, Wortart etc.) führt zu welchem Fugenelement? Der Lernalgorithmus versucht aufgrund der bereits klassifizierten Daten Regeln abzuleiten, um Voraussagen über nicht klassifizierte Komposita machen zu können. Mit korpusgrammatischem Interesse kann versucht werden, das statistisch entstandene Modell linguistisch zu deuten und mit den bestehenden Hypothesen über Regularitäten der Verfugung zu vergleichen.

Trotz dieser ersten vielversprechenden Anschauungsbeispielen über den Nutzen statistischer Verfahren besteht ein großes Desiderat, die Methoden für korpuslinguistische Zwecke anzupassen und zu verfeinern.

Ungeachtet all des Potenzials, das der korpuslinguistische Ansatz mit sich bringt, und trotz einer Vielzahl an korpuslinguistischen Spezialuntersuchungen zur deutschen Grammatik steht seine Nutzung in der wissenschaftlichen Grammatikschreibung erst am Anfang. Nicht unerwähnt bleiben dürfen allerdings die deutlichen Anzeichen in der Dudengrammatik, deren Autoren und

[2] Vgl. Gries (2008a).

Bearbeiter bei der Analyse aktueller Sprachbelege und bei der Auswahl der Beispiele auf das sehr umfangreiche Duden-Korpus zurückgreifen können.[3] Ebenfalls genannt werden muss an dieser Stelle das Projekt „Variantengrammatik des Deutschen",[4] in dem auf der Grundlage eines breiten Korpus ein Handbuch erarbeitet wird, das in alphabetisch geordneten Artikeln die areale Variation in der standarddeutschen Grammatik erfasst. Gleichwohl bedarf es für das Standarddeutsche immer noch einer korpuslinguistisch fundierten Grammatikschreibung, die der Variabilität prinzipiell jeglicher Art nachgeht und ein Werk etwa von der Art hervorbringt, wie es mit der *Longman Grammar of Spoken and Written English* (Biber et al. 2006) für das Standardenglische bereits seit einiger Zeit vorliegt.

Im Projekt „Korpusgrammatik"[5] geht es darum, diese Lücke zu schließen. Das Vorhaben bahnte sich schon seit langem an. Bereits die *Grammatik der Deutschen Sprache* (Zifonun et al. 1997) hatte bei ihrem entschieden deskriptiven Anspruch intensiv mit den Belegen aus den Korpora des IDS gearbeitet. Die empirische Orientierung verstärkte sich in den Projekten „Systematische Grammatik" und „Grammatik in Fragen und Antworten", in denen Komponenten des Grammatischen Informationssystems des IDS *grammis*[6] entstanden. In der „Grammatik in Fragen und Antworten" wurde sie sogar zum Richter in grammatischen Zweifelsfällen erhoben. Der allmähliche Übergang von den deskriptiv ausgerichteten, aber noch traditionell eher kompetenzbasierten Projekten zu dem Unternehmen, eine neue, dezidiert korpusorientierte Grammatik des Deutschen zu schreiben, wurde in jeder Phase von Bruno Strecker begleitet, der an allen genannten Projekten beteiligt war und erster Projektleiter von „Korpusgrammatik" wurde. Er ist auch in diesem Band – da er ein Kapitel übernommen hat – als Verbindungsglied zu der grammatikografischen Traditionslinie des IDS präsent.

Das Vorhaben „Korpusgrammatik" begann mit einer Orientierungsphase: Aus den vergangenen Projekten wusste man um die Bedeutung der Variation im Standarddeutschen – verschiedene Listen von besonders varianten grammatischen Bereichen und Kataloge von relevanten grammatischen Phänome-

[3] Vgl. Duden (2009: Vorwort).

[4] Unter der Leitung von Christa Dürscheid (Zürich), Stephan Elspaß (Augsburg) und Arne Ziegler (Graz) (vgl. Dürscheid et al. 2009). Dieses Projekt kooperiert mit dem hier vorzustellenden Vorhaben „Korpusgrammatik".

[5] Siehe die Homepage des Projekts unter http://www.ids-mannheim.de/gra/korpusgrammatik.html.

[6] http://www.ids-mannheim.de/grammis/.

nen lagen bereits vor.[7] Man rang aber immer noch um die Auffassung von Standarddeutsch und eine für das Vorhaben adäquate korpuslinguistische Herangehensweise. Diese Zeit des Brainstormings wurde durch die Organisation der Dritten Internationalen Konferenz „Grammar & Corpora" (2009)[8] und des Workshops „Korpora und die Grammatik des Standarddeutschen" (2010)[9] strukturiert und brachte eine Reihe von Pilotuntersuchungen[10] hervor. Es folgte eine Phase, in der Elke Donalies bereits die erste größere Studie erarbeitete, in der sie an variierenden Wortformen und Fugenelementen in zusammengesetzten Substantiven die Möglichkeiten eines korpusgrammatischen Ansatzes auslotete (Donalies 2011). Parallel dazu wurde im Projekt intensiv daran gearbeitet,

- eine für das Unternehmen „Korpusgrammatik" adäquate Korpusbasis zu erstellen,
- ein System aufzubauen, in dem das sprachliche Material vorgehalten und optimal abgefragt werden könnte, und
- eine standardisierte Methodik für statistische Analysen zu entwerfen und im Abfragesystem zu implementieren.

Das vorliegende Buch dokumentiert eben diese Bemühungen. Sie hatten insgesamt zum Ziel, notwendige konzeptuelle, empirische und methodische Voraussetzungen für das Projekt zu schaffen, – eine Grundlage, die einerseits noch weiter ausgebaut werden und auch modifizierbar bleiben muss, andererseits aber schon jetzt erlaubt, mit der eigentlichen grammatikografischen Arbeit an der Korpusgrammatik zu beginnen.

Dass diese Grundlage kein unverrückbares Fundament ist, sondern bis zum gewissen Grad flexibel bleiben muss, liegt schon an der Vagheit des Begriffs „Standardsprache". In der Forschung zur Standardsprache werden oft Ausdrücke wie *Gebrauchsstandard* oder *Gebrauchsnorm* bemüht, um darauf hinzuweisen, dass es sich bei der Standardsprache nicht unbedingt um einen kodifizierten Standard handelt, der vielleicht auch als eine Soll-Norm verstanden werden könnte, sondern um eine Konvention der Sprachgemeinschaft, die

[7] Siehe z.B. unter http://www.ids-mannheim.de/grammis/grammatikfragen/ und http://www.ids-mannheim.de/gra/korpusgrammatik.html.

[8] In Kooperation mit der Albert-Ludwigs-Universität Freiburg, vgl. Konopka et al. (2011).

[9] Mit Beteiligung der IDS-externen Kollegen Christa Dürscheid, Stephan Elspaß und Arne Ziegler (alle im Projekt „Variantengrammatik des Deutschen") sowie Anette Frank (Heidelberg), Kathrin Kunkel-Razum (Dudenredaktion) und Anke Lüdeling (Berlin).

[10] Kubczak/Konopka (2008), Strecker (2010b), Konopka (2010), Strecker (2011), Konopka (2011).

sich im Sprachgebrauch manifestiert. In der Wirtschaft spricht man in ähnlichen Fällen von einem Industrie- oder De-facto-Standard. Es gibt keine verbindliche intensionale Definition von Standarddeutsch, und wir wissen auch nicht allzu gut über seine Grenzen Bescheid, – anderenfalls wäre ein bedeutender Teil unseres Unternehmens überflüssig gewesen. Wir müssen daher den „Kandidatenkreis" für standarddeutsche grammatische Phänomeine weitgehend offen halten und unsere Methodik an die Fortschritte der Forschungsarbeit zum Wesen dieser Phänomene anpassbar machen. Wir müssen schließlich auch mit der technologischen Entwicklung Schritt halten, um letztendlich zu einer zeitgemäßen und differenzierten Beschreibung dessen zu gelangen, welche grammatischen Phänomene unter welchen Umständen als standarddeutsch gelten können, wie sie strukturiert sind und in welchen Relationen zueinander sie stehen.

Welchen Mehrwert wird die geplante Korpusgrammatik gegenüber traditionellen Forschungsansätzen erbringen?

Sie soll stark deskriptiv ausgerichtet sein und in dieser Hinsicht an die am IDS entstandene *Grammatik der Deutschen Sprache* (Zifonun et al. 1997) anknüpfen. Zurzeit ist vorgesehen, sich auch grammatiktheoretisch an diesem Werk zu orientieren, wobei das kategorialgrammatische Gerüst wie schon in der „Systematischen Grammatik" zugunsten einer oberflächennahen Konstituentenstrukturgrammatik mit syntaktischen Funktionen deutlich zurückgebaut werden soll. Zunehmen wird dagegen die Sprachgebrauchsorientierung – schon aufgrund der erst im Gebrauch manifesten Zielvarietät „Standardsprache" sowie des korpuslinguistischen Ansatzes – und die empirische Fundierung der grammatischen Aussagen wird auch eine Schlüsselposition bekommen. Im Rahmen dieses Herangehens soll es möglich werden, einerseits einen höheren Grad an Detailtreue zu erreichen, der auch auf die Aufdeckung von bisher nicht erfassten Mustern, Strukturbildungen und Regularitäten abzielt, andererseits genau die Frequenz und Distribution von Phänomenen zu beschreiben und dabei die relevanten Varianz- und Variationsfaktoren zu erfassen. All dies läuft auf eine eingehende und differenzierte Beschreibung der standardsprachlichen Grammatik hinaus, wie sie von der Forschung gefordert wird, eine Beschreibung, in der Variationsphänomene als integraler Bestandteil der Standardsprache erscheinen und in ihrem Wesen genauer erklärt werden.

**Beispiel**

Eine pauschale Feststellung wie „Die Präposition *dank* schwankt in ihrer Rektion zwischen Dativ und Genitiv“,[11] kann durch einen Panoramablick auf große Mengen standardsprachlicher Daten, wie sie Korpora bieten können, und durch statistische Analyse der auf *dank* folgenden Kontexte sehr weitgehend präzisiert werden. Schnell zeigt sich etwa, dass auf *dank* folgende Pronomina sich tendenziell anders verhalten als Nomina und dass für Letztere wiederum von Bedeutung ist, ob ihnen ein Artikel oder ein Adjektiv vorangeht oder nicht. Es kristallisiert sich das folgende Bild heraus: Anaphorische Personalpronomina stehen fast ausschließlich im Dativ (*dank ihm* und kaum *dank seiner*), Nomina mit Artikel oder Adjektiv weit überwiegend im Genitiv (*dank des Sieges*, *dank vieler Helfer* und selten *dank dem Sieg*, *dank vielen Helfern*), unbegleitete Nomina im Plural im Dativ (*dank Zuschüssen*) und im Singular ohne Kasuskennzeichnung (*dank Dosenpfand*).[12] So wird eine zuerst sehr pauschale Aussage zur Variation durch in Korpusanalysen entdeckte sprachimmanente Faktoren strukturiert und präzisiert, für bestimmte Bereiche auch stark relativiert bzw. gar unzutreffend gemacht (vgl. **dank Zuschüsse*).

Die grammatische Beschreibung wird zum beträchtlichen Teil auf der Evaluation bereits existierender Forschungsthesen zur standarddeutschen Grammatik und ihrer Variabilität beruhen. Auch hier soll an die Ergebnisse der grammatischen Forschung am IDS angeknüpft werden.

**Beispiel**

Einen guten Anknüpfungspunkt bieten etwa Bernd Wieses (2009) Untersuchungen zu starker und schwacher Adjektivflexion nach Pronominaladjektiven (z.B. *einige kleine Kinder* vs. *alle kleinen Kinder* oder *manche junge Frauen* vs. *manche alten Weiber*), in denen systematische Grundlagen der in Grammatiken beschriebenen Schwankungen geklärt werden. Dort werden zwei ineinandergreifende Parameter ausfindig gemacht, die die Verteilung starker und schwacher Formen steuern: Der abnehmenden syntaktisch-semantischen Ähnlichkeit des Pronominaladjektivs zum definiten Artikel entspreche die zunehmende Tendenz, das nachfolgende Adjektiv stark zu flektieren (in der Reihenfolge *all*, *welch*, *sämtlich*, *beid*, *solch*, *manch*, *folgend*, *mehrer*, *einig*, *viel*, *wenig*, *ander*), und die zunehmende kategorielle und formale Markiertheit habe wiederum die Zunahme der Tendenz zu Folge, das nachfolgende Adjektiv schwach zu flektieren (in der Reihenfolge N./A. Pl, G. Pl., N. Sg. Msk., D./G. Sg. Fem., N./A. Sg. Ntr., D. Sg. Msk./Ntr.).[13] Durch Korpusuntersuchungen und statistische Analysen zu Variation zwischen starker und schwacher Adjektivflexion bei verschiedenenen Kombinationen aus Pro-

[11] Dazu z.B. Duden (2007: 212), Duden (2009: 606,611), Di Meola (2009: 205f.) und weitgehend im Sinne der hier präsentierten Ausführungen Kubczak (2008).

[12] Konkrete korpuslinguistische Analysen könnten dieses Bild mit Sicherheit noch sehr weitgehend verfeinern.

[13] Siehe Wiese (2009: 190).

nominaladjektiv und Genus-Numerus-Kasus-Spezifikation könnte jetzt überprüft werden, inwiefern dieses systemlinguistisch herausgearbeitete Modell auch als Erklärung für Schwankungen im aktuellen Sprachgebrauch herhalten kann.

Von dieser Herangehensweise, die auf empirischer Aufarbeitung und statistischer Validierung von bisherigen Erkenntnissen mithilfe moderner Verfahren beruht, erhoffen wir uns auch, bisher übersehene linguistische Zusammenhänge aufzudecken und eine Neuformulierung von Regeln einzuleiten, wie es im Kapitel 5 zur Untersuchung von Fugenelementen mittels maschinellem Lernen angedeutet wird.

**Beispiel**

Bei Komposita, deren Erstglieder auf Konsonanten enden, ist das Endungssuffix dasjenige Merkmal, das die Fuge am besten voraussagt (*-keit* führt zu einer *s*-Fuge, z.B. *Geschwindigkeitskontrolle*, *-er* zu einer Nullfuge, z.B. *Siegerehrung*, etc.). Alle anderen Merkmale (phonologische und morphologische) spielen eine weit geringere Rolle. Anders bei den Komposita mit vokalisch endenden Erstgliedern: Da ist zunächst die Betontheit der letzten Silbe des Erstglieds entscheidend für die Voraussage des Fugenelements, dann, bei unbetonten Erstgliedendungen, das Flexionsparadigma, dem das Erstglied folgt. So führen z.B. Komposita, deren Erstglied mit einem Schwa endet, maskulin ist und schwach flektiert wird, zu einer *n*-Fuge (Beispiele: *Löwe-n-anteil*, *Nächste-n-liebe*). Interessant ist zudem, dass die Eigenschaften des Zweitglieds so gut wie keine Rolle zu spielen scheinen. Diese Beobachtungen resultieren aus der Analyse von über 400 000 Komposita, die nach unterschiedlichen linguistischen Eigenschaften klassifiziert wurden und auf die ein Algorithmus des maschinellen Lernens angewandt wurde, um ein Modell abzuleiten, das die Wahl des Fugenelements bei einem bestimmten Kompositum voraussagen kann.[14]

Der korpuslinguistische Ansatz eröffnet außerdem neue Perspektiven auf die Grammatik und bringt damit neue Lösungen für alte grammatische Probleme hervor, indem z.B. Kookurrenzuntersuchungen außer den erwarteten auch überraschende Kookurrenzpartner zutage bringen, die Wirksamkeit bisher unbekannter Analogiekräfte offenlegen und dadurch auch andere als traditionell-grammatische Erklärungen für bestimmte Variationsphänomene nahelegen.

**Beispiel**

Die inzwischen allenthalben sichtbare Variation zwischen *dieses Jahres* und *diesen Jahres* lässt sich mithilfe traditioneller Vorstellungen von der Systemhaftigkeit der Sprache kaum erklären. Durch Korpusrecherchen wird jedoch zum einen schnell

[14] Vgl. für die weiteren Details Kapitel 5.

ersichtlich, dass sich *Jahres* besonders häufig mit auf *-en* endenden Adjektiven ohne Artikel verbindet. Dabei sind folgende Kombinationen am augenfälligsten: *vergangenen Jahres*, *nächsten Jahres*, *kommenden Jahres* und *letzten Jahres*.[15] Zum anderen rücken bei Suchen nach Strukturen, die zu *dieses Jahres* analog sind (Artikelwort + Genitiv), Verbindungen in den Blick wie *welchen Inhalts* und *jeden Jahres*.[16] Der Weg ist somit frei einerseits für Phänomenerklärungen aus kontextuellen, inhaltlichen und ungewöhnlichen strukturellen Zusammenhängen heraus, die hier die traditionell-grammatische Paradigmatik ersetzen, andererseits für eine skalare Modellierung des schulgrammatisch gesehen stufenlosen Übergangs vom Artikel zum Adjektiv, etwa in der Art <*der*, *ein*, *dies*, *welch*, *mein*, *viel*, *jed*, *all*, *selb*, *selbig*, *folgend*, *vorig*>.[17] Auch die theoretische Frage, wie sich die Regeln der Standardsprache herleiten und welchen Charakter sie haben, kann jetzt neu gestellt werden.

Zuallerletzt ergeben sich durch die Kombination einer auf einer sehr breiten Datenbasis operierenden Empirie mit der Statistik ganz neue Interessen und Fragestellungen im Hinblick auf die Variabilität grammatischer Strukturen im Sprachgebrauch. Musterhaft und wegweisend wirken hier die Fragestellungen, die immer wieder in der Korpusgrammatik des Englischen von Biber et al. (2006) ins Blickfeld rücken und die meist die quantitativ gefasste Distribution grammatischer Erscheinungen in definierten Teilkorpora betreffen. Man vergleiche z.B. Themen aus dem Bereich Infinitivkonstruktionen („infinitive clauses", 693ff.) wie „Register[18] distribution of verb patterns" (verb + NP + *to*-clause, verb + *to*-clause etc.), „Overall frequencies of the most common verbs controlling to-clauses", „Common controlling verbs across registers".

All dies verspricht nicht nur neue Erklärungen grammatischer Zusammenhänge, sondern auch einen Anstoß für Weiter- und Neuentwicklungen im Bereich der grammatischer Theoriebildung, da die Möglichkeiten empirischer Validierung von Thesen einen großen Schritt vorangebracht werden sollen. In jedem Fall wird mit der geplanten Grammatik auch eine Anschubhilfe für neue korpusgrammatische Spezialuntersuchungen geleistet. Und schließlich deuten schon die oben angeführten Beispiele an, welch direkte Relevanz eine Korpusgrammatik des Deutschen nicht allein für die Grammatikforschung, sondern auch für den Deutschunterricht, den Unterricht von Deutsch als

[15] Vgl. Strecker (2010a).

[16] Vgl. Heringer (2011).

[17] Heringer (2011) – interessanterweise lässt sich diese Lexemreihenfolge nur zum Teil mit der weiter oben zitierten Anordnung Wieses zur abnehmenden Artikelähnlichkeit bei Pronominaladjektiven in Einklang bringen, was eine eventuelle empirische Überprüfung der Zusammenhänge auf einer breiten Datenbasis attraktiv macht.

[18] *Conversation, fiction, news, academic writing.*

Fremdsprache und *last not least* für die (ansonsten für Linguisten kaum erreichbare) sprachlich interessierte Öffentlichkeit haben könnte.

Was leistet in dem Gesamtvorhaben einer wissenschaftlichen Korpusgrammatik der vorliegende Band?

Er soll eine Momentaufnahme sein, die ziemlich genau den Stand von Januar 2012 der Bemühungen um eine Grammatik dieser Art am IDS festhält. Wer am aktuellsten Entwicklungsbild interessiert ist, kann sich im „Korpusgrammatik"-Modul von *grammis*[19] über das Fortschreiten der Arbeiten informieren. Dort werden auch die geplanten korpusgrammatischen Studien dokumentiert. Dem Momentaufnahmecharakter des Bandes entspricht auch die Tatsache, dass das Vorhaben aus ganz verschiedenen Blickwinkeln beleuchtet wird. Auf diese Weise spricht das vorliegende Werk eine recht breite Zielgruppe mit Interessen in den Bereichen Grammatik, Korpuslinguistik, Korpusaufbereitung und Korpusverwaltung an. Ein methodisch und von der Zielsetzung her homogenes Ganzes wird es erst mit dem Endprodukt, einer Korpusgrammatik, annähernd geben können. Die Einzeldarstellungen bauen dennoch aufeinander auf, und ihre Abfolge spiegelt unsere Gedankengänge wider. Kapitel 1 bewegt sich um die Grundideen, d.h. die Auseinandersetzung um das Standarddeutsche, die Auffassung von Variabilität in der Grammatik und theoretische Möglichkeiten, diese strukturiert zu behandeln, sowie um die Korpusbasis des Projekts. In Kapitel 2 findet das Ringen um ein den Ansprüchen des Vorhabens adäquates korpusgrammatisches Verwaltungs- und Abfragesystem seinen Niederschlag, das insbesondere für sehr große Sprachkorpora und eine komplexe Mehrebenenannotation geeignet ist. In Kapitel 3 werden die auf dem Abfragesystem aufsetzenden deskriptiven und induktiven statistischen Routinen gezeigt, die entwickelt wurden, um bei aller Varianz und Variation doch die Standardsprache zu fassen zu bekommen. In Kapitel 4 werden Vorteile, welche die Annotation von Korpora verspricht, seziert und einige Schwierigkeiten festgehalten, die Annotationen für Korpusanalysen mit sich bringen. Kapitel 5 soll schließlich die korpuslinguistischen Möglichkeiten veranschaulichen, abseits aller traditionell-linguistischen Verfahren mit explorativen statistischen Verfahren zu Erkenntnissen zu gelangen, die zumindest in puncto Zuverlässigkeit den heutigen Wissensstand voranbringen.

[19] http://www.ids-mannheim.de/kogra.

Der gesamte Band ist zwar in Teamarbeit entstanden, jedoch sind für jedes Kapitel bestimmte Teammitglieder in besonderer Weise verantwortlich:

Kapitel 1 – Marek Konopka
Kapitel 2 – Roman Schneider
Kapitel 3 – Noah Bubenhofer
Kapitel 4 – Bruno Strecker
Kapitel 5 – Noah Bubenhofer, Caren Brinckmann, Katrin Hein

Der Dank der Unterzeichnenden gilt vor allem den Mitautoren Bruno Strecker, Caren Brinckmann, Katrin Hein, außerdem Saskia Schmadel für ihre unermüdliche Redaktionsarbeit und ihre nützlichen Hinweise, Anna Volodina für wichtige Richtigstellungen zu unseren Vorstellungen nicht nur über die Grammatik der gesprochenen Sprache, den anonymen Gutachtern, Elke Donalies und Norbert Volz aus der Publikationsstelle des IDS für die Endredaktion und Andreas Scholz für die Erstellung der Druckvorlage.

*Noah Bubenhofer, Marek Konopka, Roman Schneider*

# 1. Datenbasis für Untersuchungen zur grammatischen Variabilität im Standarddeutschen

Im vorliegenden Kapitel sollen zunächst die Begriffe 'Standarddeutsch' und 'Variation' problematisiert sowie korpuslinguistisch handhabbar gemacht werden, bevor grundsätzlich überlegt wird, mit welchen Korpora man die grammatische Variabilität im Standarddeutschen möglichst adäquat fassen könnte. Im nächsten Schritt werden die Architekturen einiger bereits bestehender Korpusprojekte skizziert, die eine Sprache bzw. eine Varietät im Allgemeinen im Blick haben, um schließlich zu zeigen, wie die Untersuchungskorpora des Projekts aufgebaut sind.

## 1.1 Standarddeutsch? Kaum zu fassen!

Versuche, das Standarddeutsche zu definieren, sind zahlreich und fallen erwartungsgemäß je nach Betrachtungsperspektive und primärem Forschungsinteresse unterschiedlich aus (vgl. weiter unten Tabelle 1 sowie Anhang). An dieser Stelle eine neue, ultimative Definition im klassischen, intensionalen Sinn zu entwerfen, wäre daher vermessen. In diesem Vorhaben nähern wir uns der Frage „Was ist Standarddeutsch?" deshalb gewissermaßen extensional, indem wir zunächst fragen, welche grammatischen Phänomene zum Standarddeutschen gehören und wo sie sich manifestieren, um danach zu überlegen, wie ihre Domänen mithilfe von Korpora zu fassen sind.[1]

### 1.1.1 Was ist Standarddeutsch?

Im Gegensatz zu manch naiver Vorstellung ist Standarddeutsch (die Orthografie ausgenommen) nur ansatzweise explizit normiert, insbesondere gibt es kein verbindliches Regelwerk, in dem die Grammatik des Standarddeutschen festgeschrieben wäre. Sie scheint sich mehr oder weniger ungeplant zu ergeben und ist von impliziten Normen[2] durchzogen. Diese Normen bemüht man

[1] Wollte man schon die vorliegenden Ausführungen als einen Versuch verstehen, das Standarddeutsche zu definieren, so ist dagegen einzuwenden, dass hier in Umfang und Komplexität das für eine klassische linguistische Definition Zulässige gesprengt wird.

[2] Oft wird in diesem Kontext von Gebrauchsnormen gesprochen, vgl. z.B. Hennig (2009: 28), Barbour (2005: 325). Ammon (1995: 88), Barbour (2005: 325), Dürscheid et al. (2011) benutzen auch den Begriff „Gebrauchsstandard".

sich zwar in Büchern wie der Duden-Grammatik (Duden 2009)[3] zu rekonstruieren und damit zu kodifizieren. Solche Werke können dann auch zusammen mit anderen Autoritätsinstanzen[4] (z.B. Lehrern) in vielen Fragen als Orientierung dienen und ein Stück weit die Stabilität der Standardsprache fördern, aber sie wirken doch in erster Linie a posteriori und werden allem Anschein nach nur von Fachleuten ernsthaft rezipiert. Was Standarddeutsch eigentlich ist, bleibt schwer fassbar und kaum zu überblicken.

Interessanterweise können sich nichtsdestoweniger viele kompetente Benutzer des Deutschen über die standarddeutsche Grammatik unterhalten und dabei sicher sein, dass sie im Kern über dasselbe reden.[5] Das scheint nicht zuletzt daran zu liegen, dass es eine Menge grammatische Erscheinungen gibt, hinsichtlich derer Einigkeit herrscht, dass sie zum Standarddeutschen gehören (z.B. die schwache Adjektivflexion nach bestimmtem Artikel wie in *die grammatisch**en** Formen*). Bei einer weit geringeren Anzahl von Erscheinungen gehen die Meinungen auseinander bzw. ist man sich nicht sicher, ob sie dazu gehören oder nicht (z.B. die gleiche Adjektivflexion nach Pronominaladjektiv wie in *?einige grammatisch**en** Formen*). Und schließlich sind unzählige grammatische Gebilde denkbar, die man gemeinhin nicht zum Standarddeutschen zählen würde (z.B. **zahlreiche grammatisch**en** Formen*).

Standarddeutsch ist so gesehen ein reales Phänomen, das größtenteils qua unausgesprochener Konvention zustande kommt. Es wird im adäquaten Sprachgebrauch wahrgenommen und oft als konsistentes System gedacht oder zumindest als ein Inventar von Konstruktionen (eventuell einschließlich der Regeln für den adäquaten Gebrauch), das ein konsistentes System zu rekonstruieren erlaubt. Was davon aber in Wirklichkeit als greifbar übrig bleibt, ist nur der Status einzelner Erscheinungen: Zum Standarddeutschen gehört all das, was die Sprachgemeinschaft für standarddeutsch hält.[6] Mithilfe von Urteilen kompetenter Benutzer müsste man theoretisch also mehr oder weniger genau bestimmen können, welche grammatischen Erscheinungen zum Standarddeutschen gehören und welche nicht.

[3] Vgl. den dort im Vorwort erhobenen Anspruch: „Die Dudengrammatik beschreibt die geschriebene und gesprochene Standardsprache der Gegenwart."

[4] Ammon spricht dabei von (standard-)normsetzenden Instanzen (z.B. Ammon 2005: 32ff.).

[5] Dazu auch Eichinger (2005b: 143).

[6] Diese Feststellung kann nicht als eine geeignete Definition herhalten, denn sie ist nicht direkt linguistisch handhabbar.

Diese „saubere" Methode, das Standarddeutsche zu fassen, mag in Untersuchungen zur Akzeptabilität von exemplarischen Konstruktionen etwas taugen, aber um die standarddeutsche Grammatik zu *erfassen*, ist sie leider aus praktischen Gründen ungeeignet, denn wie soll man an eine aussagekräftige Menge von Urteilen zu jeder einzelnen der unzähligen grammatischen Erscheinungen im Sprachgebrauch kommen? Praktikabler als die direkte Auseinandersetzung mit dem Status einzelner grammatischer Erscheinungen ist die Betrachtung größerer Ausschnitte[7] des Sprachgebrauchs. Die Annahme dabei lautet: Wenn solche Sprachgebrauchsausschnitte von der Sprachgemeinschaft dem Standarddeutschen zugeordnet werden, dann werden auch alle darin enthaltenen grammatischen Erscheinungen in der Regel dem Standarddeutschen zuzuordnen sein. So „erbt" auf einmal eine ganze Reihe grammatischer Erscheinungen zumindest bis zu einer späteren Einzelüberprüfung die Statuszuordnung des Sprachgebrauchsausschnittes, in dem sie erscheinen. Das Feld der Kandidaten für die Bausteine der standardsprachlichen Grammatik wird deutlich verkleinert und dadurch handhabbar.

Wer mit Korpora arbeitet, hat in der Regel ohnehin nicht direkt mit grammatischen Erscheinungen zu tun, sondern mit meist digital gespeicherten Texten[8] bzw. Textfragmenten. Die Frage, ob die einzelnen Texte „Domänen des Standarddeutschen" repräsentieren, ist aber höchstens bei kleineren Textsammlungen durch Umfragen unter kompetenten Benutzern des Deutschen, die die Texte inspiziert haben, lösbar. Sobald die Texte richtig zahlreich werden, was bei Untersuchungen, die auf das Standarddeutsche in seiner Differenziertheit abzielen, naturgemäß sehr schnell passiert, muss man nach anderen Wegen suchen, um an das in der Sprachgemeinschaft vorhandene Wissen zu kommen, was zum Standarddeutschen gehört und was nicht.

### 1.1.2 Was sind standarddeutsche Texte?

Statt einzelne Texte daraufhin zu prüfen, ob sich in ihnen das Standarddeutsche manifestiert, kann man auch ganze Gruppen von Texten in Augenschein nehmen. So wie kompetente Sprachbenutzer imstande sind, einzelne Texte im

[7] Der Blick wandert so von einzelnen grammatischen Formen zu größeren sprachlichen Einheiten, in welche sie eingebettet sind. Das ist auch aus einem anderen Grund zweckmäßig, denn manche Formen können zwar isoliert als standardsprachlich bewertet werden, in falschem Kontext gebraucht vielleicht aber nicht auf alle standardsprachlich wirken, man vergleiche z.B. den e-Dativ in *am Fuße des Berges* und in *?auf freiem Fuße* oder das Präteritum in *Das Gericht hat für Recht erkannt, ...* gegenüber *?Das Gericht erkannte für Recht, ...* in der schriftlichen Fassung eines Urteils.

[8] Gesprochene Beiträge sind hier mitgemeint.

Hinblick auf die Standardsprachlichkeit zu beurteilen, bringen sie auch bestimmte Texttypen eher mit standardsprachlichem Sprachgebrauch in Verbindung als andere. Das heißt wiederum, dass bei bestimmten Arten von Texten (man denke etwa an eine wissenschaftliche Abhandlung wie diese, den Feuilletonteil einer Zeitung oder einen Lexikonartikel) von vornherein ein standardsprachlicher Sprachgebrauch erwartet wird: Es ist daher davon auszugehen, dass Autoren[9] solcher Texte diese fast immer in der Bemühung entwerfen, Standarddeutsch zu verwenden. „Fast immer", weil Autoren bisweilen auch absichtlich mit der Erwartung des Standarddeutschen spielen. Überdies muss die subjektive Überzeugung, Standarddeutsch zu verwenden, natürlich auch nicht immer voll und ganz den Tatsachen entsprechen. Derartige „Abweichungen" müssten aber als solche erkennbar sein, vorausgesetzt, man hat genug Texte des relevanten Typs betrachtet (ansonsten würde übrigens auch das Spiel mit der Erwartung des Standarddeutschen nicht funktionieren): Zahlenmäßig dürften „abweichende" Konstruktionen gegenüber ihren standardsprachlichen Varianten in einer größeren Sammlung von Texten eines Typs, bei dem Standardsprache erwartet wird, deutlich untergehen. Spätestens hier kommt eine „statistische" Denkweise ins Spiel, die auch die weiteren Überlegungen durchziehen soll.

Um Texttypen zu identifizieren, die Standarddeutsch erwarten lassen, braucht man nicht auf Umfragen unter kompetenten Sprachbenutzern zurückzugreifen, denn die bisherige Forschung zu „Standarddeutsch", „Standardsprache" und verwandten Begriffen liefert uns bereits genug explizite Kriterien für „standarddeutsche" Texte bzw. Texttypen. Man kann auch annehmen, dass die Forschungsmeinungen im Allgemeinen so etwas wie die Simulation der Urteile kompetenter Sprachbenutzer über Standardzugehörigkeit anstreben.

In Tabelle 1 wurden Merkmale[10] zusammengetragen, die in ausgewählten Definitionen und Beschreibungen von Standarddeutsch, der deutschen Standardsprache, Standardvarietät u.Ä. eine besondere Rolle zu spielen scheinen.[11] Am häufigsten betont wird die Eigenschaft 'geschrieben' (*geschriebene Spra-*

[9] Einschließlich der Sprecher gesprochener Beiträge.

[10] Die Diskussion, ob es sich dabei um hinreichende oder notwendige Bedingungen für die Standardsprache handelt, braucht hier nicht geführt zu werden – dazu siehe Dovalil (2006).

[11] Vgl. die Zusammenstellung von Zitaten im Anhang. Sie hat – wie auch Tabelle 1 – nur einen exemplarischen Charakter. Nichtdestoweniger ist davon auszugehen, dass sich bei einer Ausweitung der Literaturauswertung die Ergebnisse nicht grundsätzlich ändern würden. Sammlungen von Definitionen zu Standardsprache u.Ä. bieten auch Löffler (2005: 13ff.) und Dovalil (2006: 59-63).

*che/Erscheinungsform der Sprache, in geschriebener Form, schriftlich, schriftliche Sprachform* etc.). Damit wird aber in der Regel nicht ausgeschlossen, dass Standardsprache auch eine gesprochene Dimension hat. Die Eigenschaft 'gesprochen' ist nahezu genauso allgegenwärtig. Sie wird oft in einem Atemzug mit 'geschrieben' genannt (*in geschriebener als auch in gesprochener Form, sowohl mündlich als auch schriftlich* etc.). Jedoch ist insgesamt gesehen eine stärkere Gewichtung von 'geschrieben' zu verzeichnen. Sie wird nicht zuletzt in Konzeptionen deutlich, die aus dem englischsprachigen Raum stammen: So wird Standardsprache in Ausführungen Durrels (1999) u.a. mit *Alphabetisierung* in Verbindung gebracht und als *Sprache des gesamten Schriftverkehrs* bezeichnet, wogegen ein ähnlich gewichteter Hinweis auf ihre gesprochene Dimension fehlt; in der Definition von Haugen (1994: 4340), derer sich Elspaß (2005b: 64) in seinem Beitrag zur Entwicklung der Standardisierung in Deutschland bedient, wird auch nur eine einheitliche Schreibnorm als Voraussetzung für *standard language* genannt.[12] In anderen Beiträgen wird darauf hingewiesen, dass sich Standarddeutsch zunächst als geschriebene Sprache entwickelte (z.B. Eichinger 2005b). Jäger (1980: 377) schließlich will in seinem etwas älteren Artikel „den Terminus 'Standardsprache' der geschriebenen Sprache vorbehalten".

[12] Elspaß (2005b: 64) bemerkt dazu auch: „Dass eine Standardsprache eine Rechtlautung aufweisen muss, scheint – etwa im Vergleich mit der englischen Sprache – eine deutsche Spezialität zu sein" und verweist auf entsprechende Literatur.

| | Ammon (2005) „Standard-sprache“/ „Standard-varietäten“ | Barbour (2005) „standard language“ | Barbour/ Stevenson (1998) | Bußmann (2008) | Duden (1999) | |
|---|---|---|---|---|---|---|
| geschrieben | | X | X | X | X | |
| normiert[13] | X | X | | X | | |
| (auch) gesprochen | | X | X | X | X | |
| kodifiziert[14] | X | | X | | | |
| überregional | X | | | X | X | |
| als maßgeblich akzeptiert | | X | X | | | |
| durch Massenmedien/ Behörden/Institutionen verbreitet[15] | X | | | X | | |
| in Schulen unterrichtet | X | | X | X | | |
| variierbar | X | | | | | |
| (sozial-)prestigeträchtig | | | X | | | |
| offiziell/öffentlich/amtlich | X | | | X | | |
| von Gebildeten getragen | | X | | | | |
| nicht-mundartlich | | | | | X | |
| von Mittel- bzw. Oberschicht getragen | | | | X | | |
| (historisch) legitimiert | | | | X | | |
| allgemein verbindlich | | | | | X | |
| nicht schichten-/ gruppenspezifisch | | | | | X | |
| alle Register umfassend | | X | | | | |
| durch Selektion entstanden | | | | | | |
| funktional differenziert | | | | | | |
| mit Nation/Nationalstaat assoziiert | | | | | | |

**Tab. 1: Attribute von Standarddeutsch, Standardsprache, Standardvarietät u.Ä. in ausgewählten Beschreibungen[16]**

[13] Hierzu wurden einfachheitshalber auch Auffassungen gezählt, in denen Standard(-sprache) mit Norm gleichgesetzt wird (z.B. Barbour 2005).

[14] Hierzu wurden auch Formulierungen wie „in Grammatiken und Wörterbüchern zu finden“ (vgl. Barbour/Stevenson 1998) gezählt.

| | Durrel (1999) | Eichinger (2005a, b) | Elspaß (2005a, b) | Glinz (1980) | Jäger (1980) | Lüdtke/ Mattheier (2005) | Summe |
|---|---|---|---|---|---|---|---|
| | X | X | X | X | X | | 9 |
| | X | X | X | | X | X | 8 |
| | (X) | X | | X | | | 6 (7) |
| | X | | X | | X | X | 6 |
| | X | | X | | X | | 6 |
| | X | | X | | | | 4 |
| | X | X | | | | | 4 |
| | X | | | | | | 4 |
| | X | X | X | | | | 4 |
| | X | X | | | | | 3 |
| | X | | | | | | 3 |
| | X | | | | X | | 3 |
| | | X | | X | | | 3 |
| | | | | | X | | 2 |
| | | | | | | X | 2 |
| | | | | | | X | 2 |
| | | | | X | | | 2 |
| | | | | | | | 1 |
| | X | | | | | | 1 |
| | | X | | | | | 1 |
| | X | | | | | | 1 |

[15] Hier wurden Fälle zusammengetragen, in denen Medien, Behörden und/oder Institutionen als Träger, Vermittler und Vorbilder genannt wurden.

[16] Sofern in den Spaltenüberschriften kein anderer Terminus erscheint, handelt es sich um Beschreibungen von „Standardsprache".

Man ist natürlich versucht, unter Benutzung der in der Forschungsliteratur häufig auftretenden Attribute eine neue intensionale Definition von Standarddeutsch zu konstruieren. Bei Berücksichtigung von jenen Attributen, die in Tabelle 1 viermal oder öfter erfasst sind und dort grau schattiert erscheinen, könnte dann der erste Entwurf ungefähr so lauten:

> Standarddeutsch ist eine vorwiegend geschriebene, aber auch gesprochene Form des Deutschen, die normiert sowie (zumindest teilweise) kodifiziert ist und in der es auch Raum für Variation gibt. Sie wird überregional gebraucht, ist von der gesamten Sprachgemeinschaft als maßgeblich akzeptiert und wird durch Medien, Behörden und Institutionen verbreitet sowie in der Schule unterrichtet.[17]

Uns soll es hier jedoch um etwas anderes gehen: Es gilt zu prüfen, ob sich von den mehr oder weniger konsensuellen Attributen des Standarddeutschen Charakteristika ableiten lassen, die auf Texte zutreffen müssten, bei denen Standarddeutsch zu erwarten wäre. Tabelle 2 zeigt, dass eine ganze Reihe von Attributen des Standarddeutschen sich nur auf dessen Gebrauch bzw. Normen oder Regeln im Allgemeinen bezieht und daher kaum sinnvoll auf konkrete Äußerungseinheiten wie Texte übertragen werden kann. Aber es bleiben immer noch einige Attribute, die sich gut als Charakteristika von Texten modellieren lassen. Diese in Tabelle 2 hervorgehobenen Charakteristika kann man als Ausprägungen von den ebenfalls dort angeführten Parametern auffassen.

[17] Die Definition könnte natürlich fast beliebig weiter verfeinert werden, was hier aber nicht zweckmäßig erscheint.

| **Attribut der Standardsprache (hervorgehoben, wenn auf einen einzelnen Text übertragbar)** | **Parameter** |
|---|---|
| **geschrieben** | **Medium (Konzeption)** |
| normiert | [Bezug auf Gebrauch im Allgemeinen] |
| **gesprochen** | **Medium (Konzeption)** |
| kodifiziert | [Bezug auf Normen bzw. Regeln im Allgemeinen] |
| **überregional** | **regionale Reichweite, Region** |
| **als maßgeblich akzeptiert** | [Bezug auf Normen bzw. Regeln im Allgemeinen, betrifft auch] **regionale und soziale Reichweite** |
| **durch Massenmedien/Behörden/ Institutionen verbreitet** | **Emittent** |
| in Schulen unterrichtet | [Bezug auf Normen bzw. Regeln im Allgemeinen] |
| variierbar | [Bezug auf Gebrauch im Allgemeinen] |
| (sozial-)prestigeträchtig | [Bezug auf Gebrauch im Allgemeinen] |
| **offiziell/öffentlich/amtlich** | **Situation** |
| **von Gebildeten getragen** | **Bildungsgrad von typischem (1) Autor, (2) Adressat** |
| **von Mittel- bzw. Oberschicht getragen**[18] | **sozialer Status von typischem (1) Autor, (2) Adressat** |
| (historisch) legitimiert | [Bezug auf Normen bzw. Regeln im Allgemeinen] |
| allgemein verbindlich | [Bezug auf Normen bzw. Regeln im Allgemeinen] |
| **nicht mundartlich** | **regionale Reichweite** |
| **nicht schichten-/gruppenspezifisch** | **soziale Reichweite** |
| alle Register umfassend | [Bezug auf Gebrauch im Allgemeinen] |
| durch Selektion entstanden | [Bezug auf Normen bzw. Regeln im Allgemeinen] |
| funktional differenziert | [Bezug auf Gebrauch im Allgemeinen] |
| mit Nation/ Nationalstaat assoziiert | [Bezug auf Gebrauch im Allgemeinen] |

**Tab. 2: Attribute des Standarddeutschen und textypologische Parameter**

[18] Solche Einschränkungen werden von einigen als sozialschichtenwertende Kriterien interpretiert. Ihre Heranziehung in Tabelle 2 wie auch im Weiteren ist aber keineswegs so gemeint, sondern nur als deskriptive Wiedergabe von Meinungen anderer Forscher zu verstehen.

Basierend auf den auf Texte übertragbaren Attributen könnte man jetzt wiederum eine Definition des „standarddeutschen Texttyps" basteln wie:

> Standarddeutsche Texte sind meist geschriebene, aber auch gesprochene Texte, in denen Ausdrücke und Konstruktionen mit überregionaler Geltung vorherrschen und in welchen mundartliche und sozialgruppenspezifische Ausdrücke und Konstruktionen gemieden werden; standarddeutsche Texte sind typisch für offizielle Anlässe und öffentliche Emittenten sowie für die Kommunikation unter Gebildeten und Angehörigen der Mittel- bzw. Oberschicht.[19]

Zweckmäßiger erscheint hier aber der Versuch, die Liste der Parameter, die aus den ermittelten Charakteristika resultieren, durch genauere Unterkategorisierung operationalisierbar und auch für Übergangsphänomene tauglich zu machen wie im folgenden Entwurf, in dem Unterkategorien hervorgehoben sind, die tendenziell Standarddeutsches erwarten lassen:

A:. Medium (Konzeption):[20] z.B. gesprochen[21] (konzeptionell gesprochen), **gesprochen (konzeptionell geschrieben), geschrieben**[22] **(konzeptionell gesprochen), geschrieben (konzeptionell geschrieben)**,

B: regionale Reichweite: z.B. kleinräumig, regional,[23] **überregional, national, übernational**,

C: soziale Reichweite: z.B. spezifisch für eine bestimmte soziale Schicht/Gruppe, **nicht schichten-/gruppenspezifisch**,

D: Emittent: z.B. Privatperson, **Person in öffentlicher Funktion, (juristische) Person des Privatrechts, (juristische) Person des öffentlichen Rechts (sogenannte öffentliche Institution)**,

E: Situation: z.B. privat/ungezwungen, **öffentlich, offiziell/amtlich**,

F: Bildungsgrad von typischem (1) Autor, (2) Adressat: z.B. Grundbildung, Sekundarbildung, **Fachbildung, Hochschulbildung**,

---

[19] Ein Beispiel für einen Text, der dieser Spezifikation entspräche, ist eine Mitteilung des Rates für deutsche Rechtschreibung (z.B. „Aktualisierung des amtlichen Wörterverzeichnisses im Fremdwortbereich" vom 28.07.2011, vgl. http://rechtschreibrat.ids-mannheim.de/download/mitteilung1111.pdf; Stand: 04.01.2012), die ohne Bedenken beschrieben werden kann als 'geschrieben, überregional, weder mundartlich noch durch sozialgruppenspezifische Ausdrücke eingeschränkt, offiziell und von einer öffentlichen Institution herausgegeben sowie von Gebildeten getragen'.

[20] Vgl. z.B. Koch/Oesterreicher (2008), siehe auch Kapitel 1.4.3.

[21] In der Terminologie von Koch/Oesterreicher (z.B. 2008) „phonisch".

[22] In der Terminologie von Koch/Oesterreicher (z.B. 2008) „graphisch".

[23] Die Unterkategorie 'regional' könnte über den zusätzlichen Parameter 'Region' auch weiter spezifiziert werden, z.B. norddeutsch, mitteldeutsch, süddeutsch oder westdeutsch, ostdeutsch. Auch die Unterkategorie 'national' könnte natürlich ausdifferenziert werden.

G: sozialer Status von typischem (1) Autor, (2) Adressat: z.B. niedrig, **mittel, hoch**.

Was die einzelnen Parameter genauer zu bedeuten haben, muss noch diskutiert werden (vgl. dazu Kapitel 1.4.2). Festzuhalten bleibt zunächst, dass Texte, die das Standarddeutsche erwarten lassen, sich in etwa unter die Kategorien der oben hervorgehobenen Bereiche einordnen lassen müssten. Einem Attribut der Standardsprache entsprechen jetzt meist einige feinkörnigere Kategorien für standarddeutsche Texte. Dies bildet eine bessere Grundlage für Zuordnungsentscheidungen in Zweifelsfällen, und wird auch eher der Tatsache gerecht, dass in der Forschung im Kontext von Standard, Substandard bzw. Nonstandard oft von einem Kontinuum gesprochen wird (z.B. Löffler 2005: 21f. oder spezifischer Elspaß 2005b: 93f.).

### 1.1.3 Was sind standarddeutsche Textsorten?

Kategorisierungen wie in Kapitel 1.1.2 direkt auf einzelne Texte anzuwenden, ist in Vorhaben, die mit einer extrem breiten Korpusbasis arbeiten müssen, nicht praktikabel. Sie lassen sich aber glücklicherweise auch auf traditionelle Textsorten beziehen, wie sie im Bewusstsein der Sprachteilhaber verankert sind (wie „Enzyklopädie-Artikel" oder „Geschäftsbrief"). So sind dann auf einmal alle Texte einer als passend kategorisierten Textsorte Kandidaten für die Aufnahme in die Korpora, mit welchen man die Standardsprache analysieren will. Auf diese Weise wird das Standarddeutsche vom Phantom zu einer als solche anerkannten Existenzform der Sprache, deren Inventar an Konstruktionen einschließlich ihrer Verwendungsmöglichkeiten mittels prinzipiell durchführbarer Verfahren in Korpora zu fassen ist.

Nun stellt sich natürlich die Frage, welche konkreten Textsorten wie stark und in welchem Verhältnis zueinander in den Untersuchungskorpora vertreten sein sollten, um Rückschlüsse auf die vermutlich sehr ausdifferenzierte Standardsprache zu erlauben. Eine Antwort, die nicht zuletzt auch Machbarkeitskriterien zu berücksichtigen hat, muss aber für Kapitel 1.4 aufgespart bleiben, in dem genau die Bildung und Strukturierung der dem Projekt zugrunde liegenden Korpora beleuchtet wird. Hier sei nur noch einmal betont, dass es sich bei diesen Korpora zunächst einmal um eine Datenbasis für Analysen handelt. Erst nach diesen Analysen soll es möglich sein, zuverlässige Aussagen darüber zu machen, welche grammatischen Erscheinungen unter welchen Umständen zum Standarddeutschen gerechnet werden können, und die relevanten Erscheinungen genau zu beschreiben.

Am Ende dieses Kapitels soll noch einmal das primäre Ziel unseres Vorhabens in Erinnerung gerufen werden. Es geht darum, grammatische Formen und Formverwendungen zu untersuchen, die tendenziell zum Standarddeutschen gehören und Varianz aufweisen bzw. einer Variation unterliegen. Dies impliziert, dass etwa folgende Fragen – zumindest vorerst – zweitrangig sind: Wie ist Standarddeutsch genau zu definieren? Wie ist Standarddeutsch entstanden? Wo und in welchem Maße ist es kodifiziert? In welchem Maße wird es gelehrt? Notwendig erscheint dagegen, jetzt die Begriffe Varianz und Variation zu problematisieren.

## 1.2 Stabile Variabilität? Wie geht das?

Denkt man an die Standardisierung, scheint die Vereinheitlichung nicht weit zu sein. Der Gedanke, dass das Standarddeutsche irgendwie stabil sein muss, schwingt auch in Attributen wie 'überregional' und 'nicht schichtenspezifisch' mit. Andererseits tauchen in Tabelle 1 Charakteristika wie 'variierbar' und 'funktional differenziert' auf. Wie lässt sich dann die Idee einer Stabilität über Regionen und soziale Schichten hinweg mit den Vorstellungen von der Variabilität und der funktionalen wie auch stilistischen Ausdifferenzierung der Standardsprache vereinen?

Zum einen muss man annehmen, dass die Standardsprache in Abhängigkeit von bestimmten Aspekten der Kommunikationssituation (Anlass, Partner, Thematik, Textsorte etc.) variieren kann, und zwar in verschiedenen Regionen und sozialen Gruppen in ähnlicher Weise (vgl. Ammon 2005: 28). Zum anderen muss man aber einen gewissen Spielraum für diatopische und diastratische Variabilität zulassen. Schließlich ist auch an eine Variabilität innerhalb einer bereits genau bestimmten Kommunikationskonstellation zu denken, die z.B. durch eine konkrete Textsorte vorgegeben ist, und sei es, dass diese Variation im Sinne von *variatio delectat* oder als Vermeidung von Wiederholungen zu interpretieren wäre.

Dass die Standardsprache variiert, ist in der Forschung heute unstrittig und wird auch vielerorts betont und ausführlich diskutiert (z.B. Ammon 1995, 2005 – vor allem zur nationalen Variation, außerdem Eichinger 2005a, b; Elspaß 2005a, b). Was ist aber eigentlich mit „variieren", „Variation" u.Ä. gemeint? Angesichts zahlreicher Interpretationsmöglichkeiten sind für unser Vorhaben offensichtlich einige Sprachregelungen erforderlich. Die zu behandelnde Problematik soll dabei in ihrer Gesamtheit weiter mit dem terminologisch wenig

vorbelasteten Begriff „Variabilität" umrissen werden. In den nächsten Abschnitten werden diejenigen Ausprägungen der Variabilität konkretisiert, die das Projekt beschäftigen sollen.

### 1.2.1 Variabilität bezüglich mehrerer grammatischer Phänomene

Unabdingbar scheint die Unterscheidung zwischen dem Wechsel mehrerer sprachlicher Phänomene einerseits und Veränderungen, die nur ein Phänomen betreffen, andererseits. Nimmt man die erste Erscheinung, auf die Linguisten vorrangig mit **Variation** referieren, unter die Lupe, stößt man sogleich auf ein Paradox: Zwei oder mehr Phänomene sollen auf der einen Seite gleichwertig sein und auf der anderen Seite doch anders. Wie soll das gehen? Die entscheidende Rolle kommt hier der Bestimmung des Gemeinsamen, des Tertium Comparationis der Phänomene, die hier Vergleichsobjekte heißen sollen, zu. Alles kann an sich nämlich eine Variante von etwas anderem sein, vorausgesetzt man findet das entsprechende Tertium Comparationis. So sind auch ein Satzstrukturbaum und eine Bauzeichnung Varianten, wenn man sie unter dem Gesichtspunkt „grafische Darstellung" betrachtet. Zu einer vernünftigen Handhabe des Begriffs „Variation" kommt man daher nur, wenn man die Vergleichsobjekte und das Tertium Comparationis näher bestimmt.

In diesem Vorhaben soll es um grammatische Variation in der Standardsprache gehen. Als Vergleichsobjekte kommen projektspezifisch[24] sprachliche Einheiten ab Morphemebene[25] in Frage, die potenziell der Standardsprache angehören und sich grammatisch interpretieren lassen bzw. grammatische Unterschiede zu anderen Vergleichsobjekten aufweisen. Das Gemeinsame, das die Vergleichsobjekte erst zu Varianten macht, kann dabei sein:

(1) die Struktur und/oder Bedeutung von Syntagmen, in die die Vergleichsobjekte eingebettet erscheinen,

(2) die (text-)grammatische Funktion der Vergleichsobjekte,

(3) die grammatische Kategorisierung (das Paradigma), zu dem die Vergleichsobjekte gehören.

[24] Phonische Phänomene werden nicht behandelt, da sich das Projekt weitgehend auf geschriebene Sprache konzentriert (vgl. http://www.ids-mannheim.de/gra/korpusgrammatik.html, Stand Januar 2013), grafische Erscheinungen werden im angeschlossenen IDS-Projekt „Univerbierung" (vgl. Fiehler 2009) untersucht.

[25] Außer den typischerweise als Morpheme betrachteten Wortbildungs- und Flexionselementen (wie *be-* und *-st* in *beschreibst*) sollen zu dieser Einstiegsbene auch Fugenelemente (wie *-s-* und *-e-* in *Schweinsbraten/Schweinebraten*) zählen.

(4) Unter bestimmten Umständen schließlich kann es sich als berechtigt erweisen, die pragmatische Funktion der Vergleichsobjekte hinzuzuziehen (etwa *du* versus *Sie* als Kommunikantenpronomen in Privat- und Geschäftsbriefen).

(1) erscheint als wichtigstes Tertium Comparationis: Aus korpusbezogener Sicht wird hier die gleiche Distribution von Vergleichsobjekten bzw. der gleiche unmittelbare Kontext (Kotext) als Symptom für Variation gewertet. (2) und (3) können als logisch-grammatische Folgen bzw. Begleiterscheinungen von (1) gesehen werden, werden oben aber gesondert aufgeführt, um zu verdeutlichen, dass – eine entsprechende Annotation der Korpora vorausgesetzt – die Variationsdiagnostik auch ohne das Betrachten der Syntagmen erfolgen kann, in die die Varianten eingebettet sein können.

Ad (1): Das Syntagma, in welches die Varianten eingebettet sind, kann aus traditioneller Sicht auf verschiedenen Ebenen der Grammatik liegen (siehe Tabelle 3 unten für einige Beispiele, die aber nur die Bandbreite andeuten sollen). Der Begriff des Syntagmas ist in diesem Kontext mit dem der Konstruktion in konstruktionsgrammatischen Ansätzen[26] vergleichbar. Auch die Variable kann auf verschiedenen Ebenen liegen. Deren Ausprägungen können Fugenelemente, Flexionsendungen, (Wortbildungs-)Morpheme, Wörter, Wortgruppen, Sätze oder größere Textbausteine sein. Auf mehreren Ebenen kann es sinnvoll erscheinen, die leere Menge (Ø) ebenfalls als Variante zu betrachten.

Ad (2): Auch die gleiche (text-)grammatische Funktion der Vergleichsobjekte kann auf unterschiedlichen Ebenen angesiedelt sein, etwa

- Wortbildung, z.B. Variation von Fugenelementen,
- Satzgrammatik, z.B. Variation von Subjekten (Subjektsätze versus Subjekt-Infinitivkonstruktionen versus (pro)nominale Subjekte),
- Textgrammatik, z.B. Variation von Textbausteinen (etwa mehrere Einfachsätze versus komplexer Satz, syndetische versus asyndetische Verknüpfung von Sätzen, finale versus kausale Satzverknüpfung, verschiedene Realisierung der Konditionalität in komplexen Sätzen[27] etc.).

[26] Vgl. Fischer/Stefanowitsch (2008: 5ff.).

[27] Z.B. *Wenn/Falls es regnet, bleiben wir zu Hause.* vs. *Regnet es, bleiben wir zu Hause.*

| Ebene | | Beispiele für einbettende Syntagmen | Variable |
|---|---|---|---|
| MORPHOLOGIE | Wort (Lexem) | *Rindsbraten* vs. *Rinderbraten* (unterschiedlich verfugte Komposita)<br>*Werksgelände* vs. *Werkgelände* (verfugtes vs. nicht verfugtes Kompositum) | Fugenelement inkl. Ø |
| | Wort (Flexionsform) | [*des*] *Baum**es*** vs. [*des*] *Baum**s***,<br>[*am*] *Grab**e*** vs. [*am*] *Grab* | Flexionsendung inkl. Ø |
| | Phrase (bzw. Verbalkomplex) | *Anfang dies**en** Jahres* vs. *Anfang dies**es** Jahres* (Nominalphrase)<br>*wegen Umbau**s**/wegen Umbau/wegen **dem** Umbau* (Präpositionalphrase)<br>[*jmd.*] *sich bewerben **um**/**auf** die Stelle* (Verbalkomplex) | Flexionsendung<br>Kasusmarkierung[28] inkl. Ø<br>Präposition[29] |
| SYNTAX | Infinitivkonstruktion/ Teilsatz/Einfachsatz | *Nachmittags (**zu**) schlafen ist ungesund* (Weglassbarkeit von *zu*)<br>*Er begann jetzt (**damit**), intensiver zu trainieren* (Weglassbarkeit von Korrelaten)<br>*Der Vater hat **das Auto dem Sohn** hinterlassen.* vs. *Der Vater hat **dem Sohn das Auto** hinterlassen.*<br>***Sogar/auch/nur/selbst/einzig der Preis ist hier** in Ordnung.* vs. ***Hier ist sogar/auch/nur/selbst/einzig der Preis** in Ordnung.* | Funktionswort bzw. Ø<br>Korrelat bzw. Ø<br>Wortstellung (Wortgruppe) |
| | komplexer Satz | *Es gibt Tage, **an denen/an welchen/wo/wenn** alles schief geht.*<br>*Sie fürchtet sich davor, **überfallen zu werden.*** vs. *Sie fürchtet sich davor, **dass sie überfallen wird.***<br>*Er kam nach Hause; seine Frau war schon da.* vs. *Er kam nach Hause, **und** seine Frau war schon da.* (asyndetische vs. syndetische Satzverknüpfung)<br>***Obwohl** er arm war, war er freigiebig.* vs. *Er war arm, **trotzdem** war er freigiebig.* (Sub- vs. Koordination) | Relativsatzeinleitung<br>Teilsatz bzw. Infinitivkonstruktion<br>Konnektor bzw. Ø<br>Subjunktor vs. Adverb |
| TEXT | Text | [...] *Er kam nach Hause, **und** seine Frau war schon da.* [...] vs. [...] *Er kam nach Hause. Seine Frau war schon da.* [...]<br>(komplexer Satz vs. Einfachsätze) | Textbaustein |

**Tab. 3: Beispiele für Varianten einbettende Syntagmen auf verschiedenen Ebenen der Grammatik**[30]

[28] Hier kann man die Variable natürlich auch auf der höheren Ebene „Flexionsform" ansetzen.

[29] Hier wiederum kann man die Variable auch als „Rektion von *bewerben*" formulieren.

[30] Anna Volodina sind wir dankbar für einige hilfreiche Hinweise an dieser Stelle.

Ad (3): Das Ansetzen des Paradigmas als Tertium Comparationis für die Vergleichsobjekte ermöglicht es unter anderem, Fälle wie die folgenden zu erfassen, und zwar ohne Rücksicht auf die Syntagmen, in denen die Vergleichsobjekte direkt eingebettet sind:

- Variation von Tempusformen (z.B. Perfekt- versus Präteritumformen in verschiedenen Regionen),
- Variation von Verbmodusformen (z.B. Indikativ/Konjunktiv I/Konjunktiv II in indirekter Rede in verschiedenen Textsorten),
- Variation des Genus Verbi (z.B. Aktiv- versus Passivformen in verschiedenen Registern, etwa in literarischen Texten und in Fachtexten).

### 1.2.2 Variabilität bezüglich *eines* grammatischen Phänomens

Von der Variation im oben beschriebenen Sinn muss eine Variabilität unterschieden werden, die nur ein Phänomen betrifft. Gemeint sind damit hier die Schwankungen der Häufigkeit eines grammatischen Phänomens (relativ zur Anzahl von Textwörtern, Sätzen o.Ä.), die von Text zu Text, Texttyp zu Texttyp etc. feststellbar sind. Auf diese soll im Weiteren mit **Varianz** referiert werden.[31]

**Beispiel:**

So erscheint z.B. das Kommunikantenpronomen *du*, das allein schon aus sprachsystematischer Sicht zweifelsohne als standarddeutsch zu bezeichnen ist, in einem über COSMAS II[32] recherchierbaren Spezialkorpus zur Belletristik („DIV Belletristik des 20. und 21. Jahrhunderts: Diverse Schriftsteller") mit einer Häufigkeit von ca. 4600 Vorkommen pro eine Million Wörter. Dagegen kommt es im COSMAS-II-Archiv W, in dem alle Korpora geschriebener Sprache zusammengefasst werden und Pressetexte deutlich überwiegen, nur 155-mal pro eine Million Wörter vor. Der deutliche Unterschied ist natürlich unmittelbar einleuchtend, da direkte Rede und informell gestaltete Textpassagen in der Belletristik stark ausgeprägt sind, in Fachtexten dagegen sehr selten erscheinen.

Bei umfangreichen Korpora, die für Vorhaben wie das unsere erforderlich sind, erscheint es sinnvoll, die Varianz vorrangig auf der Ebene verschiedener Korpusteile zu betrachten bzw. – präziser – auf der Ebene von Teilkorpora, die

[31] Ob bei Varianzuntersuchungen auch die Varianz im Sinne des statistischen Streuungsmaßes angewandt oder auf ein anderes Messverfahren zurückgegriffen wird, bleibt vorerst offen. Genaueres dazu in Kapitel 3.2.

[32] Ein am IDS entwickeltes Korpusrecherche- und -analysesystem (vgl. http://www.ids-mannheim.de/cosmas2/).

verschiedene Ausprägungen eines für die Schwankungen potenziell relevanten Parameters repräsentieren wie z.B. 'Region' oder 'inhaltliche Domäne'.[33] Um die Vergleichbarkeit von gegebenenfalls unterschiedlich großen Teilkorpora untereinander zu gewährleisten, müssen die Häufigkeitswerte für die einzelnen Teilkorpora normalisiert, d.h. auf einen gemeinsamen Nenner, z.B. eine Million Wörter, gebracht werden wie schon im oben angeführten Beispiel. Eine durchgehende Betrachtung der Varianz auf Einzeltextebene erscheint hingegen kaum handhabbar. Eine bedarfsfallartige Überprüfung der Häufigkeitswerte kann dennoch für einzelne Erscheinungen erhellend sein und manche Verzerrung vermeiden helfen.

**Beispiel:**

Die Präteritalform *frug*, für die kaum einer postulieren würde, dass sie standardsprachlich sei, hatte im COSMAS-II-Archiv W in österreichischen Texten einen unerwartet hohen Anteil von 2,4% an allen Vorkommen der 1. u. 3. Person Prät. Sg. von *fragen*[34] (0,14% in bundesdeutschen und 0,06% in schweizerischen Texten). Eine genaue Überprüfung der Fundstellen ergab allerdings, dass 191 der 222 österreichischen *frug*-Belege auf ein und denselben Zeitungsredakteur zurückgehen, was das ursprüngliche Recherche-Ergebnis stark relativiert.[35]

Kommt die Sprache auf die Frequenz von Phänomenen, so rückt auch schnell die Frage in den Blick, ob man die Frequenz nicht zu Hilfe nehmen könnte, wenn es darum geht zu ermitteln, ob ein Phänomen dem Standard zugeordnet werden kann oder nicht. Diese Frage ist wohl grundsätzlich zu bejahen, wenn auch zu bedenken ist, dass die Häufigkeit, mit der grammatische Phänomene auftreten, nicht unbedingt immer im direkten Verhältnis zu ihrer Akzeptabilität steht, da sie auch von schwer greifbaren sprachimmanenten und außersprachlichen (darunter stilistischen) Faktoren abhängen kann[36] (vgl. z.B. Conrad 2010: 237). In prototypischen Fällen zumindest jedenfalls neigt man intuitiv in einer allerersten, frequenzbasierten Annäherung an die Einschätzung der Standardzugehörigkeit, wie in Tabelle 4 dargestellt, vorzugehen.

[33] Zu im Projektkorpus tatsächlich verwendeten Variabilitätsparametern vgl. Kapitel 1.4.3.

[34] Betrachtet man alle gefundenen Vorkommen der 1. u. 3. P. Prät. Sg. von *fragen* als eine aus dem tendenziell standardsprachlichen Gebrauch gezogene Stichprobe und berücksichtigt man das 95%-ige Konfidenzintervall, so ist der Wert mit 2,4% ± 0,3% anzugeben.

[35] Vgl. Konopka (2010: 34ff).

[36] Dies illustriert schon das oben angeführte Beispiel *frug*.

| Frequenz des Phänomens | → Varianz | → Zuordnungsvorschlag | Beispiele[37] (tentativ) |
|---|---|---|---|
| **hoch** | **gering** | **allgemeiner Standard** | Konnektor *aber*, starke Flexion von *dies-* wie in *dieses Jahres*, schwaches Präteritum *fragte*, Kompositum *Antragsteller* |
| **hoch nur im Teilkorpus** (z.B. regional/ national bestimmt) | **gering**[38] **im Teilkorpus** | **Teilstandard** | *parken* in bundesdeutschen und österreichischen Texten, *parkieren* in Schweizer Texten |
| **mittel** | **mittel** | **standardnah** | schwache Flexion von *dies-* wie in *diesen Jahres* |
| **niedrig** | **hoch** | **standardfern** (einschließlich Fehler) | Fugenelement *s* in *Antragssteller* |
| (hoch in einzelnen Texten/bei einem Autor) | hoch | standardfern-idiosynkratisch | starkes Präteritum *frug* |

**Tab. 4: Varianz eines prototypischen Einzelphänomens und intuitive Einschätzung der Standardzugehörigkeit**

In Tabelle 4 wird – gemäß den Überlegungen weiter oben – angenommen, dass manche standardsprachliche Phänomene in verschiedenen Teilkorpora mit ähnlichen, andere mit sehr unterschiedlichen (normalisierten) Häufigkeiten auftreten. Genauer gesagt wird davon ausgegangen, dass grammatische Phänomene, die als standardsprachlich erscheinen, zum Teil in allen Einsatzbereichen der Standardsprache (ähnlich) häufig, zum Teil in den einen verhältnismäßig häufig und in den anderen verhältnismäßig selten sein können. Diese Konzeption läuft auf einen Standard hinaus, der sich in verschiedenen Bereichen des standardsprachlichen Gebrauchs unterschiedlich gestalten kann, und trägt damit den Postulaten der bisherigen Forschung Rechnung (z.B. Ammon 2005, Dürscheid et al. 2011: 123ff. zu national/regional unterschiedlichen Standardvarietäten). Ist die Varianz eines hochfrequenten Phänomens im gesamten Korpus oder in einem Teilkorpus gering, kann es im je-

[37] Zugegebenermaßen scheinen Beispiele wie *aber, parken/parkieren, fragte/frug* nicht so sehr ein Ergebnis grammatischer Regularitäten als idiosynkratische lexikalische Phänomene zu sein. Sie werden dennoch herangezogen, da sie in ihrer Prototypenrolle für die frequenzbasierte Standardzuordnung unmittelbar einleuchtend sind. Kapitel 3.3 können konkrete Varianzmesswerte zu einigen hier präsentierten Beispielen entnommen werden.

[38] Die Varianz innerhalb der Teilkorpora muss anhand von Häufigkeitswerten für deren Teile gemessen werden.

weiligen Bereich zum Standard erklärt werden, sodass man einerseits von allgemeinem, andererseits von regionalem Standard, Textsortenstandard, medialen Standard etc. sprechen muss, wobei Letztere alle auch als „Teilstandards" zusammengefasst werden können. Wie „hochfrequent" und übrigens auch alle Frequenz- und Varianzeinschätzungen in Tabelle 4 im konkreten Fall zu quantifizieren sind, hängt vom Phänomentyp ab und spielt für diese theoretischen Überlegungen im Prinzip erst einmal keine Rolle. In jedem Fall muss aber die Häufigkeit eines jeden Standardkandidaten in seinem Geltungsbereich über die Texte bzw. Texttypen hinweg „stabil" erscheinen. Im nächsten Schritt kann man aufgrund von Frequenz- und vor allem Varianzbeobachtungen Kandidaten für standardnahe und standardferne Phänomene – sei es wieder in Bezug auf das Gesamtkorpus oder in Bezug auf die Teilkorpora – und schließlich sogar für idiosynkratische Varianten bestimmen.

Das in Tabelle 4 präsentierte Zuordnungsmodell ist jedoch bei **niedrigfrequenten** Phänomenen kaum zu handhaben, denn bei diesen ist die Varianz naturgemäß deutlich höher als bei hochfrequenten Phänomenen. Um das Ausmaß der Varianz bei einem niedrigfrequenten Phänomen richtig zu beurteilen, muss dieses in Relation zu strukturell ähnlichen Phänomenen gesetzt werden, bei denen prinzipiell eine Frequenz der gleichen Rangordnung zu erwarten ist. Da die Varianzunterschiede zwischen den einzelnen Phänomenen hier auch etwas kleiner sind (vgl. Kapitel 1.2.3), muss das Gesamtbild quasi unter einem Vergrößerungsglas betrachtet werden.

**Beispiel:**

Die Frequenz des eindeutig standarddeutschen Kompositums *Antragsteller* variiert über Teilkorpora in ähnlichem Ausmaß wie die Frequenz der umstrittenen Fügung *diesen Jahres*. Allerdings ist *Antragsteller*/*Antragssteller* im DeReKo[39] mit etwa vier Vorkommen in einer Mio. Wörter über zehnmal seltener als *dieses Jahres*/*diesen Jahres* (vgl. Kapitel 3.3). Im Bereich der Komposita wirkt *Antragsteller* dennoch hochfrequent. Es ist auch zehnmal häufiger als *Antragssteller*, dessen Varianz sich auf dem Niveau des standardfernen *frug* bewegt (vgl. ebd.).

Solche Vergleiche sind natürlich nur durchführbar, wenn es im Korpus immerhin so viele Vorkommen des Phänomens, gegebenenfalls des Phänomens und seiner Varianten, gibt, dass sie sich auf verschiedene Korpusteile (z.B. Texte, Texttypen oder Teilkorpora) – zumindest theoretisch – verteilen können und dass es bei Häufigkeitsvergleichen prinzipiell sinnvolle Differenzen geben

[39] Zum Deutschen Referenzkorpus (DeReKo) vgl. http://www.ids-mannheim.de/kl/projekte/korpora/ sowie Kapitel 1.4.1.

kann. Dies trifft auf die Phänomene im obigen Beispiel zu.[40] Eine typische Eigenschaft von Sprache ist allerdings, dass eine große Anzahl von Phänomenen extrem selten vorkommt. So zeigt Baroni (2009: 809) für den Wortschatz, dass selbst in umfangreichen Korpora der Anteil der Lexikoneinheiten, die nur einmal vorkommen, fast 50% erreichen kann. Von ähnlichen Tendenzen ist bei grammatischen Mustern wohl auszugehen (und zwar auch wenn sie nicht lexemgebunden sind), und es ist keineswegs auszuschließen, dass hier auch standardsprachliche Phänomene betroffen sind. Dieses Problem muss in der Zukunft noch genauer untersucht werden.

### 1.2.3 Perspektiven für die Betrachtung der Frequenz

Varianz und Variation können getrennt oder auch miteinander kombiniert betrachtet werden, sodass sich insgesamt folgende Perspektiven ergeben:

- Die Häufigkeit eines grammatischen Phänomens unterliegt Schwankungen (Varianz): Ein Phänomen erscheint in verschiedenen Teilkorpora unterschiedlich häufig.
- Mehrere grammatische Phänomene konkurrieren miteinander (Variation): Eine grammatische Variable hat mindestens zwei formal unterscheidbare Ausprägungen. Diesbezüglich kann in den Vordergrund rücken,
  - einerseits dass die Frequenzen der Varianten in einem (Teil-)Korpus in einem bestimmten Verhältnis zueinander stehen (frequenzbasierte Variationsbetrachtung),
  - andererseits dass sich die Varianten jeweils spezifisch auf die Teilkorpora verteilen, das heißt jeweils spezifische Varianz zeigen (varianzbasierte Variationsbetrachtung).

Bei einer **frequenzbasierten Variationsbetrachtung** kommt man schnell auf die Idee, dass die einen (im Vergleich zu den anderen häufigen) Varianten der Standardsprache zugerechnet werden könnten und die anderen (im Vergleich zu den anderen seltenen) Varianten unter Umständen nicht. Die Fragestellung wird allerdings etwas komplizierter, wenn man bedenkt, dass in verschiedenen Teilkorpora auch verschiedene Varianten überwiegen können. Die Möglichkeiten zur Einschätzung der Standardsprachlichkeit, die sich dabei modellhaft ergeben, werden für prototypische Fälle in Tabelle 5 gezeigt (zur Explikation einsetzbarer statistischer Verfahren vgl. Kapitel 3).

[40] Der seltenste der dort angeführten Ausdrücke – *frug* – war im COSMAS-II-Archiv W im Januar 2013 mit 403 Vorkommen belegt.

| Relative Frequenz[41] des Phänomens | Relative Frequenz der Variante(n) | → Zuordnungs vorschlag | Beispiele (tentativ) |
|---|---|---|---|
| **höher** | **niedriger** | **allgemeiner Standard** | *fragte* (vs. *frug*) |
| **höher** nur im Teilkorpus (z.B. regional/ national bestimmt) | **niedriger** im Teilkorpus | **Teilstandard** | *parkieren* (vs. *parken*) in schweizerischen Texten, *parken* (vs. *parkieren*) in bundesdeutschen und österreichischen Texten |
| **annähernd gleich** | **annähernd gleich** | **Standardvariante** | *gewunken* vs. *gewinkt* |
| etwas **niedriger** | etwas **höher** | **standardnah** | *diesen Jahres* (vs. *dieses Jahres*) |
| **deutlich niedriger** | **deutlich höher** | **standardfern** (Substandard, idiosynkratisch, Fehler etc.) | *frug* (vs. *fragte*), *Antragssteller* (vs. *Antragsteller*) |

**Tab. 5: Standardzugehörigkeit und Frequenz von Varianten (Tendenzen)**

In Tabelle 5 wird davon ausgegangen, dass es in einem in Teilkorpora untergliederten Korpus sowohl eine allgemeine Variation geben kann, die in allen Teilkorpora deutlich ist (z.B. *dieses Jahres* versus *diesen Jahres*, *gewinkt* versus *gewunken*), als auch eine „lokale" Variation, die nur in einzelnen Teilkorpora wirklich von Bedeutung ist (*parkieren* versus *parken* in schweizerischen Texten). Es wird außerdem angenommen, dass bei einer allgemeinen Variation das Verhältnis der Variantenfrequenzen zueinander von Teilkorpus zu Teilkorpus deutlich variieren kann (z.B. *frug* versus *fragte* in österreichischen versus bundesdeutschen und schweizerischen Texten).[42] Ist eine Variante über alle Teilkorpora hinweg im Vergleich zu anderen stark belegt, kann sie aus dieser Perspektive als allgemeiner Standard gewertet werden. Wird sie dagegen nur in einzelnen Teilkorpora häufig verwendet – immer relativ zu anderen Varianten gesehen –, so kann sie als Teilstandard gelten. Zur Erinnerung: Je nach Parameter des Teilkorpus kann es sich dabei um einen regionalen Standard, Textsortenstandard, medialen Standard etc. handeln. Sowohl in Bezug auf das Gesamtkorpus als auch in Bezug auf die einzelnen Teilkorpora kann man aus dieser Perspektive aufgrund von Frequenzverhältnissen standardnahe Varianten diagnostizieren und Kandidaten für standardferne Varianten einschließ-

[41] Die Frequenz des Phänomens im Vergleich zu seinen Varianten.

[42] Siehe Kapitel 1.2.2, vgl. auch Konopka (2010: 34ff.).

lich idiosynkratischer Varianten und Fehler isolieren. Dabei muss die Auffassung gelten, dass sich der Standard in verschiedenen Bereichen des Sprachgebrauchs unterschiedlich gestalten kann.

Gerade wenn man für bestimmte Bereiche geltenden Teilstandards auf die Spur kommen will, gewinnt **varianzbasierte Variationsbetrachtung** an Bedeutung. Denn eine verhältnismäßig hohe Gesamtfrequenz einer Variante kann z.B. über die Tatsache hinwegtäuschen, dass die Variante in bestimmten Bereichen des Sprachgebrauchs sehr häufig und in anderen sehr selten ist. Andererseits kann eine global gesehen relativ seltene Variante in bestimmten Bereichen des Sprachgebrauchs doch häufiger auftreten und dort zum Standard gehören. Und schließlich können ähnliche Frequenzen von Varianten in einem Gesamtkorpus deutliche Frequenzunterschiede in den Teilkorpora verschleiern. Die Information, ob sich eine Variante im Gesamtkorpus gleichmäßig oder nicht gleichmäßig verteilt (vgl. Kapitel 3.3 zum statistischen Instrumentarium), hilft hier weiter und leitet eventuell die genauere Untersuchung der Teilkorpora ein.

**Beispiel:**

Die Kompositumvarianten *Schweinebraten* und *Schweinsbraten* sind im DeReKo ähnlich häufig (0,47 bzw. 0,34 Vorkommen pro eine Mio. Wörter), was ohne weiteres Nachforschen suggerieren könnte, beide als gleichberechtigt dem allgemeinen Standard zuzuordnen. Allerdings sind beide Varianten auch sehr ungleichmäßig über das DeReKo verteilt (zu Messwerten vgl. Kapitel 3.3). Eine Untersuchung von Teilkorpora aus länderspezifischen Pressetexten zeigt dann folgende Lage: *Schweinsbraten* ist in österreichischen Texten deutlich häufiger als die Variante *Schweinebraten* und hier auch deutlich gleichmäßiger verteilt. In Texten aus Deutschland und der Schweiz ist dagegen *Schweinebraten* häufiger und gleichmäßiger verteilt, wobei die Unterschiede zu *Schweinsbraten* in Deutschland deutlich und in der Schweiz nur gering ausfallen (vgl. ebd.). Dies legt auf der „Länder"-Ebene in etwa folgende Zuordnungen nahe: „*Schweinsbraten* ist standardsprachlich in Österreich, eine gleichberechtigte standardsprachliche Variante in der Schweiz und eine eher standardferne Variante in Deutschland. *Schweinebraten* hingegen ist standardsprachlich in Deutschland, eine gleichberechtigte standardsprachliche Variante in der Schweiz und eine standardferne Variante in Österreich." Die Betrachtung würde natürlich genauer, wenn man von der „Länder-" auf die „Großregionen"-Ebene hinabstiege. Das Bild müsste insbesondere innerhalb Deutschlands differenzierter werden, da *Schweinsbraten* bekanntlich ein allgemein süddeutsches Phänomen ist.

Obige Ausführungen machen deutlich, dass die Dateninterpretation vielfältigen Einfluss auf die Bestimmung der Standardzugehörigkeit hat, zum einen schon durch die spezifische Zusammenstellung des Gesamtkorpus und der

Teilkorpora, zum anderen dadurch, dass es natürlich keine normierten Schwellenwerte für hohe bzw. niedrige Frequenz und für geringe bzw. hohe Varianz bei Einzelphänomenen (im Sinne von Kapitel 1.2.2) oder auch für deutliche bzw. geringfügige Frequenzunterschiede zwischen Varianten gibt. Die Arbitrarität aller Festlegungen hinsichtlich der Schwellenwerte wird besonders deutlich, wenn der Anteil des einzuschätzenden Phänomens an den Vorkommen aller relevanten Varianten im Übergangsbereich zwischen den üblichen Werten für Phänomene, welche problemlos dem Standard zugerechnet werden, und Werten typischer Nonstandard-Phänomene liegt.

**Beispiel:**

Die Verbindung *diesen Jahres*[43] hat im COSMAS-II-Archiv W einen Anteil von ca. 9% an den Verbindungen *diesen Jahres* und *dieses Jahres*. Zum Vergleich: Das standardsprachliche *gewinkt* hat einen Anteil von ca. 37% an allen *gewinkt-* und *gewunken*-Vorkommen und das nicht standardsprachliche *frug* einen Anteil von 0,37% an allen *frug-* und *fragte*-Vorkommen.[44] Bei *diesen Jahres* schwanken die Sprachbenutzer in ihrem Urteil zur Standardsprachlichkeit, was inzwischen auch schon in der Fachliteratur indirekte Spuren hinterlassen hat, vgl. z.B. „Als standardsprachlich korrekt gilt jedoch **vor allem bei konservativen Sprachpflegern** nur *Anfang dieses Jahres*" (Duden 2007: 234, Hervorhebung hinzugefügt).

Die Arbitrarität der Festlegungen für Schwellenwerte mindert aber nicht die Aussagekraft von einschlägigen Korpusuntersuchungen, wenn man sich dieser Arbitrarität bewusst ist, vorgenommene Festlegungen offenlegt, sie begründet und gemäß ihrer relativen Natur nicht allzu schwer gewichtet. Diese Arbitrarität ist das Pendant zu den Schwankungen der Sprachbenutzer bei der Variantenbeurteilung.

Unsere Hauptaufgabe sehen wir aber darin, relevante Frequenzergebnisse zu liefern, sie zu analysieren und nach Möglichkeit die Faktoren zu ermitteln, die sie steuern. Wenn dies korrekt geschehen ist, kann sich jeder, der die Projektergebnisse weiterführend interpretieren und mit Schwellenwerten arbeiten will, eigene Schwellenwerte setzen und diesen entsprechend zuverlässig einen strengeren oder lockereren Standard ermitteln.

[43] Dazu Strecker (2010) und Heringer (2011).

[44] Betrachtet man die jeweilige Gesamtheit der gefunden Varianten als eine aus dem tendenziell standardsprachlichen Gebrauch gezogene Stichprobe und berücksichtigt man das 95%-ige Konfidenzintervall, so sind die Werte mit 9,0% ± 0,1% für *diesen Jahres*, 35% ± 3% für *gewinkt* und 0,37% ± 0,03% für *frug* anzugeben.

### 1.2.4 Variabilitätsfaktoren

Für die Variabilität grammatischer Erscheinungen lassen sich bekanntlich einerseits sprachimmanente, andererseits außersprachliche Faktoren ausmachen. Betrachtet man z.B. die Fugenelemente bei Nominalkomposita,[45] so stellt sich heraus, dass das Auftreten eines bestimmten Fugenelements zum einen durch einen bestimmten unmittelbaren Kontext bedingt sein kann (z.B. die *s*-Fuge durch das Suffix *-schaft* in *Wirtschaftsminister* oder die *en*-Fuge durch das Maskulinum auf *-ent* in *Präsidentenpalast*), zum anderen durch eine bestimmte regionale Herkunft des Textautors (*Adventskalender* bei deutschen und Schweizer Autoren, *Adventkalender* bei österreichischen Autoren).

Man hätte von vornherein argumentieren können, dass bei sprachimmanenten Faktoren wie im ersten Fall mit den beiden Fugenelementen gar keine richtigen Varianten vorliegen, denn die Fugenelemente treten in einem jeweils anderen Umfeld auf. Dass ihr Auftreten eben mit einem jeweils anderen Kontext korreliert, kann man aber nicht immer im Voraus zuverlässig wissen, sondern oft erst nach einer eingehenden Korpusanalyse. Aus Gründen der Heuristik empfiehlt es sich deshalb, alle Fugenelemente vorerst (im Sinne der Bestimmungen von Kapitel 1.2.1) als Varianten mit der gleichen grammatischen Funktion als Tertium Comparationis zu betrachten. Dies scheint umso berechtigter, als in den oben angeführten Fällen sporadisch doch manchmal anders verfugt wird[46] und auf bestimmte Erstglieder mehrere verschiedene Fugenelemente folgen können (z.B. *Landesamt, Landsmann, Landtag*).

Die eine wichtige Aufgabe der geplanten Untersuchungen ist die Aufdeckung der sprachsystematischen Distributionsbedingungen für grammatische Phänomene, d. h. die Ermittlung der sprachimmanenten Faktoren ihrer Variabilität. Die andere wichtige Aufgabe richtet sich auf die außersprachlichen Faktoren. Traditionell wird davon ausgegangen, dass Sprache diachronisch, diatopisch, diastratisch und diaphasisch variiert, also in Abhängigkeit von den Parametern 'Zeit', 'Ort', 'soziale Eingruppierung der Kommunikanten' und 'Situation' (vgl. z.B. Berruto 2010: 226f.). Das Besondere an solchen außersprachlichen Parametern ist, dass man schon bei der Korpuszusammenstellung dafür sorgen kann, dass diese mutmaßlichen Variabilitätsfaktoren durch eine ent-

[45] Eine exemplarische korpuslinguistische Behandlung findet sich in Kapitel 5. Im Rahmen des Projekts „Korpusgrammatik" entstand außerdem die Pilotstudie von Donalies (2011).

[46] Man denke z.B. an *Erbschaftsteuer* (dazu Nübling/Szczepaniak 2011: 49), die Null-Fuge tritt im DeReKo selbst in Verbindungen aus *Wirtschaft* und *Minister* (*Wirtschaft(s)minister*) 298-mal auf, d.h. in 0,3% aller Fälle.

sprechende Auswahl von Texten/Textsorten adäquat repräsentiert werden. Für dieses Vorhaben bedeutet dies zuweilen einen Balanceakt, wenn man an die einschränkenden Vorgaben denkt, welche die Festlegung auf standarddeutsche Texte/Textsorten mit sich bringt (vgl. Kapitel 1.1.2). In jedem Fall müssen sich aber die Ergebnisse der Korpusanalysen so sortieren lassen, dass die Bedeutung der mutmaßlichen Variabilitätsfaktoren überprüft werden kann, was in der Praxis auf die Einrichtung von Teilkorpora hinausläuft, welche verschiedene Ausprägungen der relevanten Parameter repräsentieren.

## 1.3 Wie werden Korpora zu Sprachen oder Varietäten im Allgemeinen aufgebaut?

Das Vorhaben, ein Korpus für die Untersuchung des Standarddeutschen zusammenzustellen, kann mit Projekten verglichen werden, in denen zum Zwecke verschiedener Untersuchungen Korpora aufgebaut werden, die eine bestimmte Sprache (bzw. Varietät) im Allgemeinen repräsentieren sollen (engl. *general corpora*, in diesem Sinne wird oft auch die Bezeichnung „Referenzkorpus" gebraucht). Beispiele für in diesem Bereich mögliche und bis jetzt eingesetzte Korpusarchitekturen sollen hier drei Korpusprojekte zur englischen Sprache[47] und eines zur deutschen liefern.[48]

Zum Schlüsselproblem wird in allen diesen Projekten die Heterogenität des Untersuchungsobjekts, einer ganzen Sprache oder einer ganzen Varietät, zusammen mit der Frage, wie man diese Heterogenität so in ein Korpus hinübertransportiert, dass sie in den Korpusanalysen erkennbar bleibt und bei der Beschreibung der Ergebnisse adäquat berücksichtigt werden kann. Das dazu gehörige Stichwort lautet **Repräsentativität**.[49] Als nächstliegende Lösung für die Aufgabe, Repräsentativität zu erzielen, erscheint es, eine Art Spiegelbild

[47] In dieser exemplarischen Auswahl nicht berücksichtigte, aber grundsätzlich vergleichbare Projekte sind z.B. die *Bank of English* (vgl. http://www.mycobuild.com/about-collins-corpus.aspx, Stand: 26.10.2011) und das *Cambridge English Corpus* (http://www.cambridge.org/us/esl/catalog/subject/custom/item3637700/Cambridge-English-Corpus-Cambridge-English-Corpus/?site_locale=en_US, Stand: 26.10.2011).

[48] Zur Korpuszusammenstellung im Allgemeinen vgl. z.B. Hunston (2008), zu den Korpora geschriebener Sprache vgl. Nelson (2010), Hundt (2008), zur Korpustypologie vgl. z.B. Scherer (2006: 16-31) oder Lemnitzer/ Zinsmeister (2006: 102-113), die auch einen ausführlichen Überblick über deutschsprachige Korpora geben. Einen Überblick über bekannte Korpora weltweit gibt Xiao (2008).

[49] Eine sorgfältig reflektierte Diskussion der Repräsentativität einschließlich deren Beschränkungen liefert Hunston (2008: 161f.), Praktische Bemerkungen dazu, wie man ein möglichst repräsentatives Korpus baut, bietet z.B. Sinclair (2004).

des Sprachgebrauchs anzustreben. Da es aber schlichtweg nicht möglich ist, den Sprachgebrauch als Ganzes zu erfassen, und die Handhabbarkeit der Korpora auch deren Größe beschränken kann, muss wirkliche Repräsentativität eine Wunschvorstellung bleiben, und man sollte in diesem Zusammenhang wohl besser von der Gestaltung des Korpus als *möglichst* adäquaten Modells des Sprachgebrauchs sprechen.

Einen solchen Weg beschreiten die Autoren des **British National Corpus (BNC XML)**.[50] Eines ihrer erklärten Ziele ist es, mit dem 100 Millionen Wörter umfassenden Korpus „a microcosm of current British English in its entirety, not just of particular types" zu schaffen. Sie streben dieser Zielsetzung entsprechend prinzipiell an,

- sowohl geschriebene als auch gesprochene Sprache einzubeziehen,
- die geschriebene Sprache sowohl aus der Perspektive der Textrezeption als auch der Textproduktion zu betrachten und schließlich
- bei der geschriebenen Sprache sowohl Publikationen in Form von Büchern, Periodika u.Ä. als auch kleinere Publikationsformen und nichtpublizierte Texte zu berücksichtigen.

Pragmatische Gründe führen allerdings dazu, dass in allen drei Punkten die jeweils leichter aufzuarbeitenden und zu beschaffenden Dimensionen des Sprachgebrauchs im BNC das Übergewicht bekommen. So macht die geschriebensprachliche Komponente 90% des Gesamtkorpus aus, und bei dieser überwiegen Publikationen der etablierten Form, also Bücher, Periodika u.Ä., welche vor allem für die rezipierte Sprache repräsentativ sind, kaum aber für die Gesamtheit der produzierten Sprache stehen können (vgl. Abbildung 1).

[50] Informationen zum BNC wurden den Benutzerhinweisen auf der Seite http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html (Stand: 23.08.2011) entnommen, vgl. auch Leech/Rayson/Wilson (2001: 2ff.).

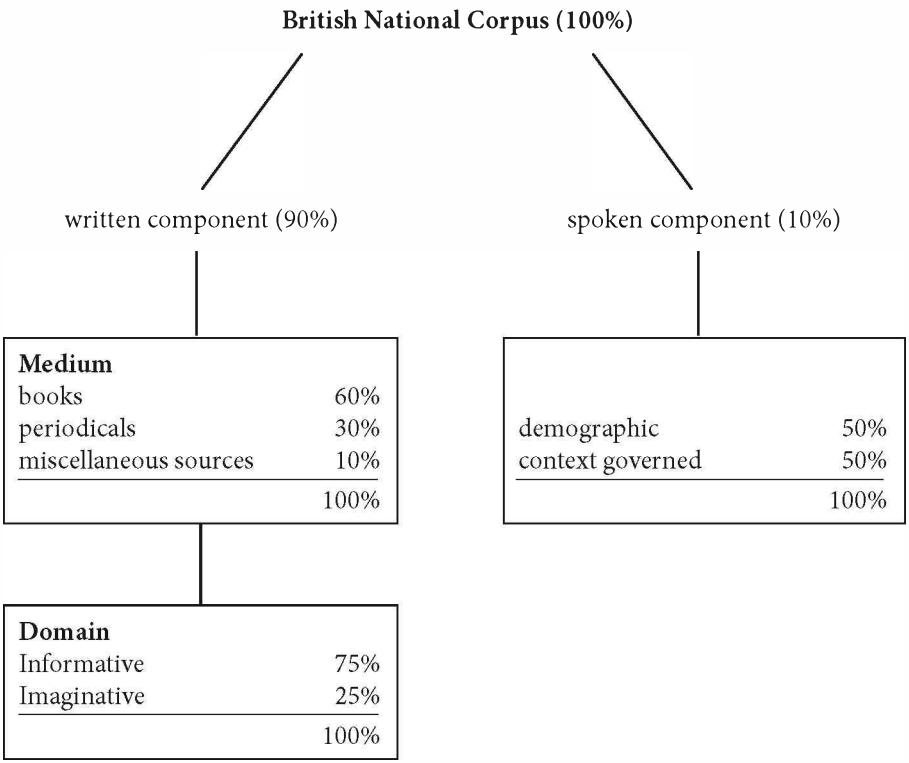



**Abb. 1: Design des BNC XML (wie unter http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html vom 23.08.2011 beschrieben)**

Die Größe einzelner Texte der geschriebensprachlichen Komponente des BNC XML überschreitet 45 000 Wörter nicht. Um eine ausreichende sprachliche Differenzierung dieses Teilkorpus zu erreichen, werden die Texte entlang zweier unabhängiger Parameter 'Medium' und 'Domäne' (inhaltlicher Bereich) ausgewählt. Die prozentuellen Zielvorgaben für die dabei entstehenden Klassen sind der Abbildung 1 zu entnehmen. Sie beziehen sich immer nur auf die Gesamtkomponente und gelten nicht innerhalb der Klassen, welche die anderen Auswahlparameter auffächern. Die Domäne *informative* wird zusätzlich in acht Subdomänen unterteilt (*natural & pure science, applied science, social science, world affairs, commerce & finance, arts, belief & thought, leisure*). Ein drittes und letztes Auswahlkriterium bildet die 'Zeit' der Textentstehung – ab 1960 bei der imaginativen (belletristischen) und ab 1975 bei der informativen Domäne.

Neben den sich durch die Auswahlkriterien ergebenden Klassifizierungen kann es im BNC auch Zusatzinformationen zu Korpustexten geben, welche, ohne dass sie die Textauswahl bestimmen, Untersuchungen zu feinkörniger bestimmten Textgruppen erlauben können. Dazu gehören Angaben zu Autor (Geschlecht, Alter, Wohnsitz etc.), Zielgruppe, Publikationsort und Textzuschnitt (ganzer Text, Anfangsfragment etc.), wobei solche Informationen bei vielen Texten auch fehlen.

Die Auswahl der Bücher richtet sich teils nach deren Verbreitung und Popularität, indem Bestsellerlisten, Verzeichnisse preisgekrönter Titel, Ausleihstatistiken sowie Auflagezahlen ausgewertet werden, um die Rezeption zu quantifizieren, teils – mehr oder weniger zufallsgesteuert – nach einem Verzeichnis lieferbarer Bücher („Whitaker's Books in Print"). Bei den Periodika sorgt man für eine breite regionale Streuung. Die sonstigen geschriebenen Quellen (*miscellaneous sources*) bestehen aus kleineren Publikationsformen (z.B. Werbezetteln), diversen nicht publizierten Texten (z.B. Essays von Schülern und Studenten) und Vorlesevorlagen (z.B. Fernsehnachrichten).

In die gesprochensprachliche Komponente des BNC wurde zum einen die spontane Sprache von nach demografischen Diversifikationskriterien wie Alter, Geschlecht, soziale Gruppe ausgewählten Sprechern aufgenommen, zum anderen monologische (40%) und dialogische (60%) Beiträge verschiedener Typen aus vier gleich stark berücksichtigten Kontextbereichen (*educational, business, public/institutional, leisure*).

Insgesamt wird beim BNC XML die Bemühung um eine Modellierung des gesamten Sprachgebrauchs deutlich, bei der faktisch aber die breit rezipierte geschriebene Sprache ins Zentrum rückt. Gesprochene und geschriebene Sprache, einzelne mediale Textklassen sowie einzelne Domänen werden im Korpus jeweils unterschiedlich stark gewichtet, was – abgesehen von pragmatischen Zwängen – teils mit der kulturellen Bedeutung, teils mit dem Publikationsvolumen der Texte begründet wird, die in den Korpuskomponenten zu repräsentieren sind. Man muss sich allerdings dessen bewusst sein, dass die Zielvorgaben für einzelne Komponenten wie 60% für 'Bücher', 30% für 'Periodika' etc. – auch wenn sie Intuitionen der Korpusautoren in Bezug auf den Einfluss der entsprechenden Textgruppen auf die Sprachgemeinschaft spiegeln mögen – letztlich arbiträr sind: Sie hätten bei anderen Korpusdesignern auch anders ausfallen können. Wer die im BNC XML vorgenommenen Gewichtungen nicht bedenkenfrei hinnehmen kann, hat aber dazu, das Gesamtkorpus als Modell des Sprachgebrauchs zu betrachten, immerhin eine interes-

sante Alternative: Da die Komponenten von BNC XML ausreichend groß erscheinen, kann man es auch als eine Sammlung von Teilkorpora betrachten, die untereinander vergleichbar gemacht werden können, indem die Häufigkeitsangaben für die einzelnen Teilkorpora mittels Normalisierung in Relation zur Größe des jeweiligen Teilkorpus gesetzt werden.[51]

Die oben angesprochene Arbitrarität der Festlegungen zu Proportionen zwischen den Teilkorpora ist charakteristisch für die Versuche, den gesamten Sprachgebrauch zu modellieren. Konzentriert man sich stärker auf die Vergleichbarkeit wie auch immer bestimmter Ausschnitte des Sprachgebrauchs untereinander, so erscheint es am günstigsten, wenn sie im Gesamtkorpus von vornherein gleich stark berücksichtigt sind.[52] Das Stichwort dazu lautet **Ausgewogenheit**. Der Begriff wird zwar unterschiedlich und teilweise sehr vage gehandhabt (dazu Sinclair 2004, Kapitel 5), aber zumindest ein sehr häufiges Verständnis davon bezieht sich auf das Gleichgewicht (bzw. die Vermeidung deutlicher Größenunterschiede) zwischen den Teilkorpora (vgl. Hunston 2008: 163, Clancy 2010: 86f.).

Die BNC-Designer haben auch diese Perspektive verfolgt. Das Ergebnis sind zwei auf BNC XML basierende kleinere Korpora:

- der **BNC Sampler**, der aus einem geschriebenen und einem gesprochenem Teil mit jeweils ungefähr einer Million Wörter besteht und ansonsten das BNC XML in seinem Design spiegelt, dabei aber eine detailliertere und manuell überprüfte Wortarten-Annotation aufweist, sowie
- das **BNC Baby**, das aus vier Teilen mit ungefähr einer Million Wörter besteht, die jeweils ein Genre repräsentieren (*academic writing, imaginative writing, newspaper texts, spontanous conversation*) und aus dem BNC XML ohne Revisionen herauskopiert wurden.

[51] Vgl. Hunston (2008: 162) zur Unausgewogenheit zwischen der gesprochensprachlichen und geschriebensprachlichen Komponente: „The spoken component comprises only 10% of the whole, which is clearly not representative either of production or of reception, but which is explained by the heavy resources required to collect spoken data in electronic form. However, 10% of 100 million words is a corpus of a respectable size that allows research to be carried out into spoken British English. The process of normalisation is used to allow valid comparisons between the written and the spoken components (Leech/Rayson/Wilson 2001). The problem of a lack of representativeness disappears."

[52] Ein einleuchtender Grund ist, dass seltene Phänomene, die in größeren Teilkorpora spärlich, aber immerhin vertreten sind, aus den kleineren Teilkorpora aufgrund ihrer Seltenheit „verschwinden" können, was aber natürlich kein Beweis dafür sein kann, dass sie in dem durch das kleinere Teilkorpus repräsentierten Texttyp nicht existent sind (das sogenannte 'Problem der negativen Evidenz').

Die Idee der Ausgewogenheit zwischen Korpusteilen bestimmt auch die restlichen hier vorzustellenden General-Corpus-Projekte zum Englischen sowie die Korpusbasis des Projekts „Digitales Wörterbuch der deutschen Sprache" (DWDS): Die regionalen Unterschiede stehen im Vordergrund des Vorhabens **International Corpus of English (ICE)**, in dem primär das Ziel verfolgt wird, Material für komparative Studien zum gesprochenen und geschriebenen Englisch weltweit zu sammeln. Dieses Ziel begründet ein Korpusdesign, bei dem die Ausgewogenheit in einer spezifischen Weise und sehr weitgehend verfolgt wird. Im ICE werden zahlreiche Korpora zu den einzelnen nationalen bzw. regionalen Varietäten des Englischen zusammengestellt, wobei jedes etwa eine Million Wörter umfasst, die in 500 Texten/Textfragmenten zu je ca. 2 000 Wörtern organisiert sind, und in Bezug auf die Texttypen zwar nicht direkt ausgewogen, aber dafür wie die anderen Korpora des ICE strukturiert ist (vgl. http://ice-corpora.net/ice/design.htm, Stand: 29.08.2011). Das ICE als Ganzes erscheint somit ausgewogen in Bezug auf den Parameter 'nationale bzw. regionale Varietät', und es stehen in allen Einzelkorpora auch dann gleiche Mengen an Texten und Wörtern zur Verfügung, wenn es darum geht, die Sprache innerhalb einer bestimmten Textklasse (z.B. 'geschriebene Sprache', 'Ungedrucktes', 'Briefe' oder 'Geschäftsbriefe') zu untersuchen.

Das Hauptziel des **LSWE Corpus**, das mit 40 Mio. Wörtern den Analysen der *Longman Grammar of Spoken and Written English* zugrunde liegt, lautet wiederum „to provide a systematic representation of different registers, particular focusing on the four registers of conversation, fiction, news and academic prose" (Biber et al. 1998: 24). Folgerichtig wird eine Ausgewogenheit im Hinblick auf den Parameter 'Register' angestrebt. Die Teilkorpora zu den vier Kernregistern umfassen zwischen 3,9 und 5,4 Mio. Wörter. Die Auswahl der Kernregister wird zum einen durch die Breite der Sprachgebrauchsabdeckung, zum anderen durch die Ökonomie von Analyse und Präsentation begründet: „The four core registers cover much of the range of variation in English, while being restricted to a manageable number of distinctions" (ebd.). Innerhalb der Register wird eine Streuung im Hinblick auf eine Reihe von weiteren Parametern angestrebt wie Geschlecht und Alter des Sprechers (*conversation*), nationale Varietät (*fiction*), Region, Thema/Ressort (*news*), Disziplin (*academic prose*), wobei es bei diesen Unterteilungen weniger um die Ausgewogenheit als vielmehr um eine Ausdifferenzierung bzw. um die Abdeckungsbreite geht. Das Korpus soll als Ganzes Gegenwartsenglisch repräsentieren und umfasst daher mehrheitlich Texte ab 1980, im Register *fiction* wegen deren tendenziell

langer Wirkung mehrheitlich Texte ab 1950. Die Textlänge variiert von durchschnittlich 250 Wörtern bei (britischen) Zeitungsartikeln bis durchschnittlich 35 000 Wörter bei fiktionalen Texten, wobei im letzteren Fall sowohl komplette Bücher als auch Teile von Büchern aufgenommen wurden. Die Kernregister werden durch zwei Zusatzregister ergänzt – *non-conversational speech* und *general prose* –, die mit ca. sechs bzw. sieben Mio. Wörtern das Korpus für Gesamtanalysen stärken bzw. bei wenigen Spezialanalysen herangezogen werden. Schließlich werden für Dialektvergleiche auch zwei Spezialkorpora zum amerikanischen Englisch (*conversation*, *news*) verwendet, wobei die Autoren betonen, dass der Dialektvergleich in der *Longman Grammar of Spoken and Written English* im Hintergrund stehe, da für grammatische Aspekte Registerunterschiede viel wichtiger seien als Dialektunterschiede (ebd.: 26).

Eine weitgehende Ausgewogenheit, und zwar im Hinblick auf die beiden Parameter 'Dekade' und 'Textsorte', zeigt schließlich das Referenzkorpus (Kernkorpus) des lexikografischen Projekts **DWDS**[53], „das als ausgewogen und hinreichend groß in Bezug auf den Forschungsgegenstand 'deutscher Wortschatz des 20. Jahrhunderts' bewertet werden kann" (http://www.dwds.de/resource/kerncorpus/, Stand: 29.08.2011):[54] Es umfasst 100 Mio. Wörter, von denen 95% annähernd gleichmäßig auf die einzelnen Dekaden und die vier Textsorten Belletristik, Zeitung, Wissenschaft und Gebrauchsliteratur (jeweils zu 20-27%) verteilt sind. Die restlichen 5% entfallen ergänzend auf transkribierte Texte gesprochener Sprache. Dem Referenzkorpus steht interessanterweise ein Ergänzungskorpus zur Seite, das „opportunistisch" angelegt ist, d.h. weniger auf Ausgewogenheit abzielt als auf Umfang und Aktualität unter Berücksichtigung leicht zugänglicher Texte.[55] Es soll für viele statistische Zwecke und für selten belegte Wörter genutzt werden (Klein 2004: 287), besteht im Wesentlichen aus Zeitungstexten ab 1980 und umfasst über 1,5 Milliarden Wörter (http://de.wikipedia.org/wiki/DWDS, Stand: 31.8.2011).

Bei allen oben präsentierten Korpusarchitekturen wurden als Zielpopulationen ursprünglich die Sprachen/Varietäten im Allgemeinen anvisiert. Die obigen Beispiele zeigen dabei, dass immer wieder drei Aspekte des Korpusdesigns in den Vordergrund rücken: die (intendierte) Abdeckung des Sprachgebrauchs

[53] Zum Projekt *Digitales Wörterbuch der Deutschen Sprache* (DWDS) vgl. Klein (2004).

[54] Genaueres zur Motivation und zum Aufbau des DWDS-Korpus in Geyken (2007).

[55] Ein mehr oder weniger „opportunistisches" Korpus, das laufend erweitert wird (sogenanntes Monitorkorpus), liegt auch der *Bank of English* (BoE) zugrunde (vgl. http://www.mycobuild.com/about-collins-corpus.aspx, Stand: Januar 2013).

(die tatsächliche Grundgesamtheit), die Proportionen zwischen den Teilkorpora und der Umfang der Gesamt- und Teilkorpora. Je nach Korpusbestimmung bzw. Interessenlage der Korpusdesigner können diese Aspekte spezifisch gestaltet werden. In Bezug auf die beiden ersten kommen dabei prinzipiell folgende schablonenhafte Lösungen vor:

- Abdeckung: (1) der gesamte Sprachgebrauch,[56] (2) durch einen Parameter (z.B. 'Register') strukturierte Textgruppen, (3) „opportunistisch" ausgewählte Sprachgebrauchsbereiche,
- Proportionen zwischen Teilkorpora: (1) (nach Möglichkeit) dem Sprachgebrauch entsprechend bzw. nachempfunden skaliert, (2) ausgewogen, (3) „opportunistisch" skaliert.

Was den Umfang des Gesamtkorpus angeht, so ist er bei dem Bezugsgegenstand (statistisch gesehen: der Zielpopulation) eine ganze Sprache bzw. Varietät tendenziell sehr groß, nichtsdestoweniger begegnen in den obigen Beispielen auch relativ kleine (Spezial-)Korpora wie BNC Sampler mit zwei Mio. Wörtern.

Der Begriff der Repräsentativität liegt gewissermaßen quer zu den oben genannten drei zentralen Aspekten des Korpusdesigns. Er lässt sich aber – je nach seiner genaueren Interpretation[57] – mit bestimmten Lösungen in Verbindung bringen. Naheliegend ist für ein möglichst „repräsentatives" Korpus einer ganzen Sprache/Varietät etwa die Konstellation: „(tendenziell) den gesamten Sprachgebrauch abdeckend", „(nach Möglichkeit) dem Sprachgebrauch entsprechend skaliert" und „möglichst groß". Allerdings sind das Zielvorgaben, die – wie schon das BNC XML zeigt – eigentlich nie endgültig befriedigt werden können. Es sind unterschiedliche Reaktionen darauf vorstellbar: Man kann (1) trotzdem versuchen, die Zielvorgaben *soweit aktuell möglich* zu erfüllen, und darin einen weiteren wissenschaftlichen Fortschritt sehen, (2) sich mehreren „wichtigen" Sprachgebrauchsausschnitten widmen, ohne eine Repräsentativität für die Gesamtvarietät zu postulieren, und schließlich (3) – zumindest bei bestimmten Fragestellungen – „opportunistisch" vorgehen und auf Differenziertheit des Materials und schiere Korpusgröße setzen in der Hoffnung, dass sie manche „Schieflage" ausgleichen. Die präsentierten Bei-

[56] Gemeint ist hier, dass der gesamte Sprachgebrauch gespiegelt bzw. modelliert und nicht im Korpus berücksichtigt wird.

[57] Zu Verschiedenheit der Repräsentativitätsauffassungen vgl. z.B. Sinclair (2004).

spiele für verschiedene Korpusprojekte zeigen, wie die Entscheidung für einen konkreten Weg bei der Korpuszusammenstellung jeweils mit einer spezifischen Konstellation von Lösungen hinsichtlich der drei zentralen Aspekte des Korpusdesigns einhergeht (siehe Tabelle 6).

| Lösung | Abdeckung | Proportionen | Umfang | Korpusbeispiel |
|---|---|---|---|---|
| Korpus als Modell des Sprachgebrauchs | der gesamte Sprachgebrauch | teilweise dem Sprachgebrauch entsprechend skaliert[58] | großes Gesamtkorpus | BNC XML |
| Korpus „wichtiger" Sprachgebrauchs-ausschnitte | durch einen Parameter strukturierte Textgruppen | ausgewogen | ausreichend große Teilkorpora | BNC Sampler<br>BNC Baby<br>ICE<br>LSWE Corpus<br>DWDS-Referenzkorpus |
| „opportunistisches" Korpus | „opportunistisch" ausgewählte Sprachgebrauchs-bereiche | zufällig skaliert | extrem großes Gesamtkorpus | DWDS-Ergänzungskorpus |

**Tab. 6: Lösungen für das Repräsentativitätsproblem bei der Korpuszusammenstellung**

Neben den Überlegungen, wie man sich mit dem Korpus am besten der Zielsprache/-varietät (z.B. „modernes Englisch" oder „Deutsch des 20. Jahrhunderts") nähert, kann schließlich auch der Charakter der Untersuchungen, die mithilfe des Korpus durchgeführt werden sollen, dessen Design beeinflussen. Während BNC und ICE prinzipiell nicht auf Untersuchungen aus einem bestimmten linguistischen Bereich ausgerichtet sind, wurden das LSWE Corpus für grammatische Untersuchungen und das DWDS-Referenzkorpus für Wortschatzuntersuchungen entworfen. Auch beim vorliegenden Vorhaben ist genau zu überlegen, auf was die Korpusanalysen zum Standarddeutschen eigentlich abzielen sollen: Sie sollen eine Beschreibung von grammatischen Phänomenen ermöglichen, bei der – und das ist der Clou – ihre Variabilität im Zentrum steht. Für das Korpusdesign relevant erscheint etwa, dass

[58] Die Komposition der geschriebensprachlichen Komponente des BNC XML wird durch die Bemühung bestimmt, sowohl rezeptions- als auch produktionsrelevante Texte gebührend zu berücksichtigen, und sich innerhalb dieser beiden Kategorien, solange es nicht auf Kosten der Differenziertheit der Texttypen geht, an der Popularität und Verbreitung der Texte zu orientieren (vgl. http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#wrides).

- die Variabilität ausfindig gemacht werden muss,
- bei Variabilität, insbesondere bei Variation, prinzipiell auch mit sehr seltenen Phänomenen gerechnet werden muss,
- nach Variabilitätsfaktoren gesucht werden soll,
- die Wirkung bereits in der Forschung anerkannter Variabilitätsfaktoren wie 'Medium', 'Region', 'Textsorte' etc. überprüft werden muss.

Das Idealkorpus müsste folglich einerseits groß und möglichst differenziert, andererseits nach Variabilitätsfaktoren strukturierbar und im Hinblick auf die bereits bekannten Variabilitätsfaktoren ausgewogen sein – insgesamt eine sehr komplexe Vorgabe, die in der Praxis schwer zu erfüllen ist. Ein großes, möglichst differenziertes Korpus macht es schwierig, eine Ausgewogenheit in Bezug auf die Variabilitätsfaktoren zu erreichen, und ein sorgfältig ausgewogenes Korpus kann hingegen nicht allzu viele Teilkorpora umfassen und wird für die Suche nach seltenen Phänomenen womöglich nicht groß genug sein.

Der in diesem Vorhaben gewählte Ausweg aus diesem Dilemma, besteht darin, die beiden Zielsetzungen voneinander zu trennen und in zwei getrennten Korpora zu verfolgen: einem Korpus, das auf Größe und Heterogenität ausgerichtet ist, und einem, das auf die Ausgewogenheit der Teilkorpora setzt, die sich nach erkannten wichtigen Variabilitätsfaktoren richten. Wie dies in der Praxis vonstattengeht, zeigt das nächste Kapitel.

## 1.4 Die Korpusbasis des Projekts

Die Repräsentativität unseres Korpus im Sinne eines Spiegelbilds des Sprachgebrauchs anzustreben, erscheint problematisch, da prinzipiell unklar ist, in welchem Verhältnis zueinander die anzusetzenden Teilkorpora stehen müssten. Eine wie auch immer geartete Lösung in diesem Bereich stichhaltig zu begründen, ist nicht nur sehr umständlich, sondern die Begründung wäre auf Dauer mit Sicherheit auch nur schwer haltbar. Im vorliegenden Vorhaben soll es darüber hinaus nicht um das Standarddeutsche an sich gehen, sondern primär um die grammatische Variabilität in diesem. Der Repräsentativitätsaspekt ist hier also hauptsächlich auf die Differenziertheit des Sprachgebrauchs zu beziehen. Wichtig ist daher vor allem, dass im primären Untersuchungskorpus (im Weiteren: „Gesamtkorpus") verschiedenartige Bereiche des Standarddeutschen gebührend berücksichtigt werden, damit ein Spektrum an Variation sichtbar wird, und weniger, dass die „realen" Proportionen zwischen diesen Bereichen

eingehalten werden. Um Varianz und Variation aufzuspüren, brauchen diese Bereiche im Gesamtkorpus nicht durch gleich große Teilkorpora vertreten zu sein. Die Teilkorpora sollten nur – jedes für sich – groß genug sein, um auch seltene Erscheinungen auffindbar zu machen und stichhaltige Aussagen zu den untersuchten Erscheinungen im Teilkorpus zuzulassen. Diese Ansprüche führen selbstverständlich zu einem sehr umfangreichen Gesamtkorpus. Eine ähnliche Alternative zu den klassischen „repräsentativen" Korpora sieht übrigens Hunston (2008: 162), die folgende Möglichkeit erwägt:

> [...] to seek to include texts from as many different sources as possible in the corpus but to treat the resulting corpus as a collection of sub-corpora rather than as a single entity. This is feasible only when each sub-corpus is of a reasonable size.

Um Häufigkeitsangaben zu einer Erscheinung in unterschiedlich großen Teilkorpora miteinander vergleichbar zu machen, können die Angaben normalisiert, also durch Hochrechnungen auf einen Nenner gebracht werden, was inzwischen auch das allgemein übliche Verfahren ist (z.B. Leech et al. 2001, Biber et al. 2006). Außerdem kann mit geeigneten Auswertungsmethoden wie etwa Signifikanztests der unterschiedlichen Größe der Teilkorpora Rechnung getragen werden. Trotz alledem erscheint es angesichts der aktuellen korpuslinguistischen Praxis doch sinnvoll, dem Gesamtkorpus ein kleineres, sekundäres Untersuchungskorpus zur Seite zu stellen, das Ausgewogenheit anstrebt und (tendenziell) gleich große Teilkorpora umfasst (im Weiteren: „ausgewogenes Korpus"). Damit werden für die Nutzer die Vergleiche zwischen den Teilkorpora unmittelbar aussagekräftig und statistische Stichhaltigkeit direkt erfahrbar. Im ausgewogenen Korpus soll man u.a. gezielt der Bedeutung der mutmaßlichen Variabilitätsparameter nachgehen können, sodass seine Teilkorpora nach diesen Parametern auszurichten sind. Mit dieser anvisierten Benutzung mehrerer Korpora geht unser Vorhaben einen Weg, der bereits in den Projekten BNC und DWDS gewählt wurde (wenn auch mit teilweise anderem Hintergrund, siehe Kapitel 1.3).

Im vorliegenden Unternehmen wurde bei der Bildung der Untersuchungskorpora zunächst auf IDS-eigene Ressourcen, vor allem auf das Deutsche Referenzkorpus (DeReKo), zurückgegriffen (künftige Ergänzungen aus anderen Quellen behalten wir uns vor). Die Korpusbasis des Projekts besteht zurzeit aus

- dem **Gesamtkorpus**, das auf DeReKo-Texten basiert, die nach 1955 (bzw. – im Bereich 'Publikumspresse' – nach 1990) und vor 2011 entstanden sind (ca. 4,3 Mrd. Token bzw. 16 Mio. Texte), und

- dem **ausgewogenen Korpus**, das im Hinblick auf die Parameter 'Medium', 'Register', 'Region' und '(inhaltliche) Domäne' soweit möglich ausgewogen ist und einen Ausschnitt des Gesamtkorpus darstellt (ca. 20 Mio. Token bzw. 20 000 Texte).

Aus praktischen Gründen wurde bei der Tokenzählung der Connexor Machinese Phrase Tagger benutzt, eines der Werkzeuge, mit denen die Korpora annotiert sind.[59] Token sind hier folglich die durch den Tagger identifizierten Wortformen und Satzzeichen. Die geschilderte Zweiteilung der Korpusbasis erweist sich insofern als zweckmäßig, als dass das Gesamtkorpus durch seinen Umfang insbesondere für die vorbereitenden Recherchen, Analysen zu grammatischen Variabilitätsbedingungen und Untersuchungen zu seltenen Phänomenen geeignet ist, während das ausgewogene Korpus eine systematische Überprüfung der grammatikexternen Distributions- und der Variationsparameter (z.B. 'Medium', 'Register'/'Textsorte', 'Region') erlaubt. Der folgenden genaueren Darstellung der beiden Korpora ist vorauszuschicken, dass deren bisherige interne Struktur nicht jedem Anspruch Genüge leistet, der in der Konzeptionsphase erhoben wurde. Dabei ist aber zweierlei zu bedenken: Einerseits liegen die Korpora in ihrer ersten „Ausbaustufe" vor und sie werden noch weiterentwickelt, andererseits bleibt eine Korpuszusammenstellung, um mit Nelson (2010: 60) zu sprechen, immer „a compromise between the hoped for and the achievable". Und schließlich muss auch angemerkt werden, dass beide Untersuchungskorpora zwar auf größtenteils öffentlich zugänglichen[60] DeReKo-Daten basieren, aber aus rechtlichen Gründen zurzeit nur projektintern zur Verfügung stehen.

### 1.4.1 Gesamtkorpus

Das DeReKo „bildet mit über 5,4 Milliarden Wörtern (Stand 29.02.2012) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit". Die Subkorpora „enthalten belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten". Sie „werden im Hinblick auf Umfang, Variabilität, Qualität und Aktualität akquiriert" und

[59] Zur Aufbereitung der Untersuchungskorpora vgl. Kapitel 2.1.

[60] Recherchierbar über das Korpusrecherche- und -analysesystem COSMAS (siehe http://www.ids-mannheim.de/cosmas2/).

enthalten ausschließlich urheberrechtlich abgesichertes Material (vgl. http://www.ids-mannheim.de/kl/projekte/korpora/).

Manche mögen das DeReKo nicht als Referenzkorpus betrachten, sondern als ein Archiv.[61] Wie dem auch sei, laut DeReKo-Team können daraus prinzipiell weitere Korpora komponiert werden, „die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sind" (ebd.). Es scheint auch durch seine Ausrichtung auf Größe und Ausdifferenzierung und nicht zuletzt durch seine urheberrechtliche Absicherung bestens als Grundlage für das Gesamtkorpus unseres Vorhabens geeignet.[62] Die Entscheidung, welche Teile vom DeReKo in dieses Korpus eingehen sollen, muss sich jetzt an den allgemeinen Zielen des Projekts und an den in Kapitel 1.1.2 erarbeiteten, für standarddeutsche Texte relevanten Parametern orientieren sowie Machbarkeitsargumente beachten.

### 1.4.2 Texttypologische Eignung von DeReKo-Texten

Das Projekt soll sich mit Standardsprache in aktuell-synchroner Hinsicht befassen. Daher ist es erforderlich, zunächst zu überlegen, wie weit die Entstehungszeit der berücksichtigten Texte zurückgehen darf. DeReKo-Texte belegen mit sinkender Dichte mehr oder weniger gut die Zeit bis ins 18. Jh. zurück. Der für die aktuelle Standardsprache relevante **Zeitraum** dürfte aber nicht so früh beginnen. Im LSWE Corpus, das für eine 1999 zum ersten Mal erschienene Grammatik des Englischen zusammengestellt wurde, stammen die Texte mehrheitlich aus der Zeit nach 1980, nur die Texte im Register 'fiction' stammen mehrheitlich aus der Zeit nach 1950 und einige „Klassiker" aus der ersten Hälfte des 20. Jahrhunderts. (Biber et al. 1999: 26ff.). Dass für die belletristischen Texte ein deutlich früheres Datum als für andere Texte gesetzt wurde, wird mit dem längeren „shelf-life" literarischer Werke begründet. Vor diesem Hintergrund erscheint es vertretbar, den Erscheinungsbeginn beim Gesamtkorpus unseres Vorhabens für Texte der populären Publikumspresse auf 1990 und für alle anderen Texte (hauptsächlich Buchbeiträge, aber auch Beiträge in Fachzeitschriften sowie kleinere Formen wie Gebrauchsanleitungen etc., vgl. Kapitel 1.4.3) auf 1955[63] zu setzen. Der genaue Zeitpunkt für die letzte Grenze

[61] So etwa Hundt (2008: 181) zu einer früheren Version von DeReKo.

[62] Grundsätzliches zum Design vom DeReKo im Kontext von Begriffen wie „Repräsentativität", „Ausgewogenheit" oder „Referenzkorpus" in Kupietz et al. (2010) sowie Perkuhn et al. (2012: 45ff.).

[63] Bei nicht journalistischen Texten wurde das Entstehungsdatum des Textes in vielen Fällen „manuell" hinzugefügt bzw. das im DeReKo verzeichnete Publikationsdatum überprüft und eventuell durch das Ersterscheinungsdatum ersetzt.

ist nicht zuletzt dadurch motiviert, dass das DeReKo für die Jahre direkt vor 1955 besonders viele Werke von Thomas Mann enthält, sodass eine Entscheidung für ein früheres Datum ein Ungleichgewicht zugunsten des idiosynkratischen Sprachgebrauchs eines einzelnen Autors hätte mit sich bringen können. Der Zeitraum, in dem die im Gesamtkorpus berücksichtigten Texte erschienen sind, endet mit dem Jahr 2010.

Im DeReko sind nur Korpora geschriebener Sprache enthalten, und dies muss in Bezug auf das vorliegende Projekt zunächst selbstverständlich als eine mediale Einschränkung gewertet werden. Allerdings ist hier daran zu erinnern, dass für die Standardsprache – wie in Kapitel 1.1.2 ausgeführt – öfter eine Priorität der Schriftlichkeit postuliert wird. Darüber hinaus ist zu bedenken, dass bei gesprochener Sprache der Übergang zwischen Standard und Nonstandard-Varietäten fließender zu sein scheint als bei geschriebener Sprache, sodass die Suche nach zum Projektgegenstand passenden gesprochenen Beiträgen schon an sich eine sehr schwierige Aufgabe darstellen würde. Hinzu kommt, dass die spezifischen Produktionsbedingungen spontan gesprochener Sprache zu einer mehr oder weniger eigenständigen „Online-Syntax" führen (vgl. Auer 2000, dazu Dürscheid et al. 2011: 125), die so deutliche Unterschiede zur Syntax geschriebener Sprache aufweist, dass sie auch anders behandelt werden müsste.[64] All dies bedeutet nicht, dass die gesprochene Standardsprache gänzlich aus dem Blickwinkel des Projekts verschwindet. Sie rückt lediglich etwas in den Hintergrund, und ihre Grammatik kann unter Umständen zu einem späteren Zeitpunkt genau aufgearbeitet werden, zumal sie einen spezifischen, gut abtrennbaren und im standardsprachlichen Gefüge nicht prioritären Teilgegenstand darstellt. Eine Zusammenstellung eines zusätzlichen, nicht auf DeReKo beruhenden gesprochensprachlichen Korpus wäre hier eine notwendige Voraussetzung. Gewissermaßen einen Ausblick auf die Mündlichkeit wird man dennoch schon mit DeReKo erarbeiten können, da darin viele Texte zu finden sind, denen konzeptionell gesprochene Beiträge zugrunde liegen wie etwa Protokolle von Bundestagsdebatten, Verschriftungen von Interviews und Gesprächen etc. Einen Trost bietet schließlich auch ein Blick auf den texttypologischen Parameter **Medium/Konzeption**, wie er in Kapitel 1.1.2 unterkategorisiert wird: Von insgesamt vier dort angesetzten Textkategorien sind im DeReKo zwar zwei nicht vertreten, aber nur eine davon – 'medial gesprochene und konzeptionell geschriebene Texte' – lässt in besonderer Weise standard-

[64] In Duden (2009) erhält die gesprochene Sprache ein separates Kapitel, das sich auch in der Beschreibungssprache deutlich vom Rest der Duden-Grammatik absetzt.

sprachliche Texte erwarten. Die andere – 'medial und konzeptionell gesprochene Texte' – ist hingegen in hohem Maße von Nichtstandardsprachlichkeit geprägt.

Der nächste texttypologische Parameter aus Kapitel 1.1.2, **räumliche Reichweite**, ist primär auf die Geltung der in den Texten enthaltenen Phänomene zu beziehen. Allein – wie lässt sich diese ermitteln? Wohl kaum, bevor man die Texte nicht untersucht hat. Damit wären aber Ergebnisse späterer Korpusanalysen, wenn nicht vorweggenommen, dann zumindest in unzulässiger Weise vorbestimmt. Nicht umsonst besteht Sinclair (2004) darauf, Korpora nur anhand textexterner Parameter zu erstellen. Die räumliche Geltung der in einem Text enthaltenen Phänomene kann allerdings grob als Funktion der anvisierten räumlichen Verbreitung des Textes behandelt werden. Bei der überwiegenden Mehrheit der DeReKo-Texte scheint es diesbezüglich keinerlei Einschränkungen zu geben, was prinzipiell Standardsprachlichkeit erwarten lässt. Im Übrigen können regional vertriebene Texte (etwa in regionalen Zeitungen) ebenfalls von Phänomenen mit überregionaler Geltung geprägt sein. Solche Texte sind auch nützlich, wenn es um die Untersuchungen zu regionalen Schattierungen in der Standardsprache geht. Zu konstatieren bleibt schließlich, dass Texte, bei denen offensichtlich eine sehr kleinräumige Verbreitung anvisiert wird (was aber keineswegs mit einem standardfernen Sprachgebrauch zusammengehen muss) im DeReKo in jedem Fall sehr selten sind.

Ähnliches ist zum Parameter **soziale Reichweite** zu sagen: Auch diese lässt sich primär auf die Geltung der sprachlichen Phänomene beziehen, die aber nicht im Voraus untersucht werden können; weicht man auf die anvisierte soziale Verbreitung der Texte als Indikator aus, so ist bei DeReKo-Texten kaum von intendierten Einschränkungen auf einzelne Gesellschaftsschichten auszugehen, wohl aber punktuell von einer Adressierung an Gruppen von Fachleuten (etwa beim Fachsprachenkorpus zur Gentechnologie (dkg)). Derartige gruppensprachliche Einschränkungen sind dokumentiert, und sie betreffen nur verhältnismäßig wenige Texte. Daher sind auch grammatische Daten aus diesen Texten isolierbar und leicht zu kontrollieren. Dadurch können sie das Bild der standardsprachlichen Grammatik auch nicht verzerren, wohl aber dazu genutzt werden, festzustellen, ob die jeweiligen Fachsprachen grammatisch gesehen von der Standardsprache abweichen.

Für den Parameter **Textemittent** spielt es eine entscheidende Rolle, dass es sich bei nahezu allen DeReKo-Texten um offiziell verlegte und tendenziell sorgfältig redigierte Druckerzeugnisse handelt. Somit ist aus dem Spektrum der Unterkategorien, die in Kapitel 1.1.2 diesem Parameter zugeschrieben wurden, lediglich die Kategorie 'Privatperson' nicht vertreten, eben die einzige, mit der in verstärktem Maße Nichtstandardsprachlichkeit verbunden werden kann.

In Bezug auf den nächsten Parameter **Situation** ist festzustellen, dass es sich bei DeReKo-Texten aus dem oben genannten Grund und auch, weil die Texte meist in höheren Auflagen erschienen sind, um eindeutig öffentliche und offizielle bzw. amtliche Konstellationen handelt. Eben mit diesen Konstellationen wird üblicherweise Standardsprachlichkeit in Verbindung gebracht. Hier spielen standardferne private bzw. ungezwungene Kommunikationssituationen kaum eine Rolle.

Was den Parameter **Bildung** des Textautors angeht, so lässt sich der Bildungsgrad zum einen bei vielen der im DeReKo erfassten Texte nicht feststellen (viele Autoren, insbesondere von Pressetexten, sind z.B. häufig einfach nicht mehr ermittelbar) und zum anderen bei dieser Menge von Texten überhaupt nur mit erheblichen Aufwand zuverlässig herausfinden, sodass hier der Lösungsweg über intuitive Pauschalurteile unvermeidbar ist. Da es sich bei allen DeReKo-Texten um offiziell verlegte und redigierte Druckerzeugnisse handelt, ist davon auszugehen, dass die Textautoren bzw. -bearbeiter in weit überwiegender Mehrheit über eine mittlere und höhere Bildung verfügen bzw. – negativ formuliert – dass Texte von Autoren, die lediglich über Grundbildung verfügen, im DeReKo ein Randphänomen darstellen dürften. Ähnliches ist zur Bildung des anvisierten Textadressaten festzustellen mit dem Unterschied, dass hier öfter von keinerlei Vorannahmen und kaum von Einschränkungen der Autoren bezüglich der Bildung der Adressaten auszugehen ist und dass dadurch das Publikum mit Grundbildung einfach mit adressiert erscheint.

**Sozialer Status von Autor und Adressat** ist bei DeReKo-Texten schließlich ähnlich zu beurteilen, allerdings mit dem Unterschied, dass sich der soziale Status schon aufgrund seiner komplexen Zusammensetzung weniger direkt als der Bildungsgrad mit der Wahrscheinlichkeit korrelieren lässt, als Autor oder Leser eines in einer größeren Anzahl von Exemplaren verlegten Druckerzeugnisses aufzutreten. Auf alle Fälle sind im DeReKo nur im geringen Umfang Texte zu erwarten, die von Autoren mit feststellbar niedrigem sozialem Status

geschrieben wurden oder die ausschließlich an so einzustufende Leser gerichtet sind.

Insgesamt ist also festzustellen, dass im Hinblick auf die in Kapitel 1.1.2 erarbeiteten texttypologischen Parameter alle Texttypen (dort Unterkategorien), die sich tendenziell mit Standarddeutsch verbinden lassen, im DeReKo vertreten oder zumindest zu erwarten sind – mit einer Ausnahme: Texte, die medial gesprochen, aber konzeptionell geschrieben sind. Mithilfe des DeReKo lassen sich damit selbstverständlich keine direkten Untersuchungen zur gesprochenen Sprache durchführen. Dieser Bereich müsste getrennt, anhand eines zusätzlichen Korpus aufgearbeitet werden. Andererseits scheinen im DeReKo keine Texte vorhanden zu sein, die in der Typenübersicht in Kapitel 1.1.2 eher mit der Nichtstandardsprachlichkeit in Verbindung gebracht wurden. Alles in allem ein äußerst zufriedenstellendes Ergebnis: Da die Größe des Gesamtkorpus für die Untersuchungen an sich nicht beschränkt sein muss, sondern im Gegenteil einen besonderen Vorteil darstellt, reicht hier die zeitliche Anpassung des DeReKo, und es müssen im Hinblick auf die Texttypologien keine Teilkorpora bzw. Textgruppen ausgeschlossen werden, um zum Gesamtkorpus für die Untersuchungen zu gelangen. Dieses Gesamtkorpus umfasst nach den am Anfang dieses Kapitels genannten zeitlichen Einschränkungen immer noch über vier Milliarden Wörter. Die extreme Korpusgröße macht es unwahrscheinlich, dass einzelne besonders lange Texte die Korpusanalysen verfälschen können. Auf diese Weise können im Untersuchungskorpus sowohl ganze Texte als auch Textfragmente berücksichtigt werden, so wie sie im DeReKo vorgehalten werden. Wie bei Sinclair (2004)[65] wird hier davon ausgegangen, dass es an sich ungünstig ist, Korpora nur auf Textfragmenten aufzubauen, denn sie können insofern nicht für die ganzen Texte stehen, als die Präferenzen für Konstruktionen von Textposition zu Textposition variieren. Zu der Praxis, Texte auf Fragmente gleicher Größe zu reduzieren, schreibt Sinclair:

> There is no virtue from a linguistic point of view in selecting samples all of the same size. True, this was the convention in some of the early corpora, and it has been perpetuated in later corpora with a view to simplifying aspects of contrastive research. Apart from this very specialied consideration, it is difficult to justify the continuation of the practice. The integrity and representativeness of complete artefacts is far more important than the difficulty of reconciling texts of different dimensions. [...] **Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete**

[65] Zum Textzuschnitt z.B. auch Hunston (2008: 165f.), Clancy (2010: 85).

> **speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.** (Sinclair 2004: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm, Stand: 2.7.2003, Hervorhebung im Original)

Es bleibt zu prüfen, ob sich innerhalb des zeitlich eingeschränkten DeReKo ausreichend große Teilkorpora für mögliche Variabilitätsfaktoren bilden lassen und wenn ja, später bei den grammatischen Analysen zu testen, ob auch innerhalb dieser Teilkorpora einzelne besonders lange Texte keine Verfälschungen verursachen.

### 1.4.3 Variabilitätsfaktoren

Im DeReKo stehen zu den einzelnen Texten folgende relevante Metadaten zur Verfügung, wobei nicht alle zu jedem Text vorliegen (vgl. genauer Kapitel 2.2.1):

- Genre, Ressort, Topic, Staat, Titel, Autorenname, Texttyp, Publikationsdatum, Publikationsort.

Ausgehend von diesen Metadaten und auch anhand weiterführender Recherchen zu einzelnen Texten konnten folgende Zusatzinformationen nachgeführt werden:

- Medium, Register, Domäne, Region, Geburts- bzw. Heimatort des Autors, Entstehungszeit (Einzelheiten in Kapitel 2.2.1).

Alle diese Daten machen es möglich, das Gesamtkorpus in verschiedenartige Textgruppen aufzuteilen, was bei Recherchen spontan immer wieder genutzt werden kann und soll. Einige Daten werden bereits von vornherein als Ausprägungen mutmaßlicher Variabilitätsfaktoren behandelt und zu Unterteilungen des Gesamtkorpus in Teilkorpora benutzt. Die Unterteilungen richten sich nach:

- Medium,
- Register,
- Staat und Region,
- (inhaltlicher) Domäne.

Dies sind durchgehend Kriterien, die bereits in zahlreichen anderen Projekten als Variabilitätsfaktoren behandelt und für die Bildung von Teilkorpora genutzt wurden. Ein weiterer Variabilitätsfaktor, der in diesem Vorhaben regel-

mäßig überprüft wird, ist die (Entstehungs-)Zeit,[66] die aber als Kontinuum nicht für die Bildung von starren Teilkorpora benutzt wird, sondern flexibler gehandhabt werden soll. Mit allen diesen Kriterien sind drei Dimensionen abgedeckt, die in der Forschung immer wieder als für die sprachliche Variation relevant behandelt werden: Zeit, Raum und Situation (zu Letzterer 'Medium' und 'Register'). Nicht direkt repräsentiert ist dagegen die diastratische Dimension.[67] Wie die Unterteilung in Teilkorpora genau aussieht, wird in Tabellen 7 bis 13 gezeigt und diskutiert.

### Medium

| Medium | Token[68] | Anteil | Texte | Anteil |
|---|---|---|---|---|
| Publikumspresse | 4 230 974 187 | 97,97% | 15 704 646 | 98,54% |
| Bücher | 9 568 628 | 0,22% | 947 | 0,01% |
| Sonstige Printmedien | 13 173 311 | 0,31% | 31 902 | 0,20% |
| Internet/Wikipedia | 60 683 615 | 1,41% | 196 732 | 1,23% |
| Gesprochenes | 4 387 933 | 0,10% | 2 853 | 0,02% |
| Gesamt | 4 318 787 674 | 100,00% | 15 937 080 | 100,00% |

Tab. 7: Unterteilung des Gesamtkorpus im Hinblick auf 'Medium'

Die Unterteilung nach 'Medium' gab es bereits im BNC XML (vgl. Kapitel 1.3). Hier erhält sie eine spezifische Ausprägung, bei der auch der Begriff „Konzeption"[69] eine Rolle spielt. Die Unterteilung lässt sich nachvollziehen, indem die Teilfaktoren 'Konzeption', 'medialer Träger' und 'Erscheinungsart' nacheinander abgearbeitet werden (im Strukturbaum in Abbildung 2 von oben nach unten).

[66] Von 1955 bzw. – bei Publikumspresse – von 1990 bis einschließlich 2010 vgl. Kapitel 4.1.

[67] Die Standardsprache an sich kann auch als bereits eingeschränkt bezüglich dieser Dimension gesehen werden, wenn man Auffassungen folgt, die sie als typisch für Gebildete bzw. für bestimmte soziale Schichten betrachten (vgl. Kapitel 1.1.2).

[68] Durch den Connexor-Tagger identifizierte Wortformen und Satzzeichen. (http://www.connexor.eu/technology/machinese/index.html, Stand: Januar 2013). Vgl. Kapitel 1.4 sowie zur Aufbereitung der Untersuchungskorpora Kapitel 2.1.

[69] Koch/Oesterreicher (z.B. 2008)

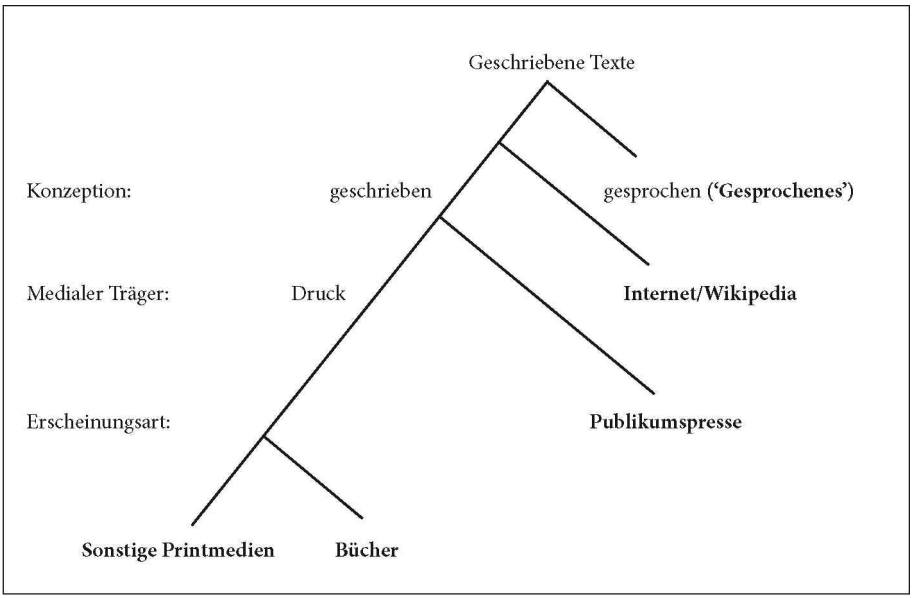



**Abb. 2: Mediale Strukturierung des Untersuchungskorpus**

Die im konzeptionellen Kontinuum stärker **gesprochenen** Texte ('Gesprochenes') bilden ein Kontrollkorpus, das gewissermaßen einen Ausblick auf die Mündlichkeit erlaubt. Die Texte gehören den im Folgenden aufgelisteten Textsorten der DeReKo-Metadatenbank an und wurden größtenteils „von Hand" auf ihre Eignung überprüft: *Ansprache, Diskussion, Rede, Gespräch, Interview, Predigt, Vortrag, Protokoll,*[70] *Rundfunkbeitrag, Fernsehsendung, Hörspiel.* Es handelt sich somit einerseits um Texte, die geschrieben wurden, um gesprochen zu werden,[71] andererseits um Verschriftungen von gesprochenen Kommunikationsbeiträgen,[72] in beiden Fällen um Textsorten aus der

[70] Z.B. „stenografische Protokolle" von Bundestags- und Volkskammersitzungen.

[71] Dies setzt konzeptionell relevante Prozesse der „Vermündlichung" (Koch/Oesterreicher 2008: 200) voraus. Zu der Gruppe können folgende DeReKo-Textsorten gezählt werden: *Ansprache, Predigt, Rede, Vortrag, Hörspiel, Rundfunkbeitrag, Fernsehsendung.* Die Texte könnten gemäß dem Ansatz Kochs/Oesterreichers (ebd. bzw. 1985: 17f.) auch als „Verschriftung" von Äußerungsformen betrachtet werden, die medial phonisch sind, konzeptionell aber (tendenziell) geschrieben. Für wertvolle Hinweise danken wir Anna Volodina.

[72] Hierzu sind die DeReKo-Textsorten *Diskussion, Gespräch, Interview, Protokoll* zu zählen. Im Ansatz Kochs/Oesterreichers wären die der Verschriftung zugrunde liegenden Äußerungen als medial gesehen phonisch und konzeptionell (tendenziell) gesprochen einzuschätzen. Hier ist auch mit konzeptionell relevanten Prozessen der „Verschriftlichung" zu rechnen (vgl. Koch/Oesterreicher 2008).

Übergangszone zwischen konzeptioneller Schriftlichkeit und konzeptioneller Mündlichkeit.

Das geplante Teilkorpus **Internettexte** beschränkt sich zurzeit aufgrund von Schwierigkeiten bei der Beschaffung urheberrechtlich abgesicherten Materials nur auf Wikipedia-Artikel (Stand 2005). Diese stellen zwar einerseits einen sehr spezialisierten Ausschnitt im Internet erscheinender Texte dar, sie bilden aber andererseits eine für das Internet typische Instanz des kooperativen Schreibens und lassen gleichzeitig eine hohe Standardsprachlichkeitserwartung aufkommen. Nichtsdestoweniger sollen in der Zukunft weitere Internet-Textsorten dazukommen.

Das Teilkorpus **Publikumspresse** enthält Tageszeitungen, populäre wöchentliche Nachrichtenmagazine und Publikumszeitschriften (vgl. Übersicht weiter unten). Ein Pendant dazu im Bereich der Printmedien bilden die Teilkorpora **Bücher** und **Sonstige Printmedien**. Ersteres enthält längere (zumindest intendiert) selbstständige Werke oder deren Fragmente. Das Kriterium 'Buch' wurde dabei relativ streng gehandhabt (zu für Bücher relevanten Merkmalen vgl. z.B. http://de.wikipedia.org/wiki/Buch, Stand: 03.01.2012), das heißt, dass im Zweifelsfall, insbesondere bei kürzeren und nicht selbständig erschienenen Beiträgen, die Entscheidung für das Teilkorpus 'Sonstige Printmedien' fiel. Letzteres wurde so zu einem Sammelbecken der Druckerzeugnisse, die weder der Publikumspresse angehören noch als Buch bezeichnet werden können. Es enthält einerseits Texte wie Aufsätze in Fachbüchern und Fachzeitschriften, Essays, Gesetzentwürfe, andererseits aber auch Gebrauchsanleitungen, Packungsbeilagen, Briefe, Prospekte, Werbung und Flugblätter. Zu überlegen wäre, ob in den nächsten Korpusausbaustufen beim Materialzuwachs in bestimmten Bereichen diese nicht in selbständige Teilkorpora wie etwa 'Graue Literatur'[73] ausgegliedert werden könnten.

In Tabelle 7 fällt das extreme Übergewicht der 'Publikumspresse' auf. Hierzu ist zunächst anzumerken, dass es in der Forschung Stimmen gibt, die behaupten, dass eben diese Textgruppe bereits allein für die empirische Fundierung der Aussagen zur Standardsprache geeignet sei (z.B. Eisenberg 2007: 217, Dürscheid et al. 2011: 126). Vor allem ist aber zu betonen, dass die Größe der anderen Teilkorpora des Gesamtkorpus jeweils für sich genommen immer noch ausreichend erscheint, um als Grundlage für aussagekräftige Analysen zur

[73] Nicht über den Buchhandel vertriebene Publikationen.

Grammatik im jeweiligen Bereich zu dienen.[74] So kann eine erste Vergleichbarkeit zwischen den Teilkorpora durch Normalisierung der Frequenzergebnisse geschaffen werden, während genauere Vergleiche ohnehin auf das sekundäre, ausgewogene Korpus auszulagern sind.

### Register

| Register | Token | Anteil | Texte | Anteil |
|---|---|---|---|---|
| Pressetexte | 4 241 649 343 | 98,21% | 15 735 407 | 98,73% |
| Gebrauchstexte | 69 598 590 | 1,61% | 201 244 | 1,26% |
| Literarische Texte | 7 539 741 | 0,17% | 429 | < 0,01% |
| Gesamt | 4 318 787 674 | 100,00% | 15 937 080 | 100,00% |

Tab. 8: Unterteilung des Gesamtkorpus im Hinblick auf 'Register'

Die Bezeichnung „Register" ist hier im Sinne von Biber/Conrad (2009: 6) gemeint:

> In general terms, a ***register*** is a variety associated with a particular situation of use (including particular communicative purposes). The description of a register covers three major components: the situational context, the linguistic features, and the functional relationships between the two components [...].[75]

Damit sind (tendenziell) größere Textsorten- bzw. Genregruppen gemeint, die sich situativ definieren und hinsichtlich funktional motivierter, typischer (d.h. im Text besonders häufiger) linguistischer Charakteristika zu beschreiben sind. Biber/Conrad kontrastieren Register mit Genre, indem sie Letzteres nicht mit häufigen linguistischen Charakteristika, sondern mit konventionellen textstrukturierenden Merkmalen verbinden. Während bei Genre-Untersuchungen immer ganze Texte zu betrachten seien, kann die Register-Perspektive sowohl auf der Betrachtung von ganzen Texten als auch Textfragmenten

[74] Biber (1993) wandte mathematische Verfahren an, um zu berechnen, wie groß das Korpus sein muss, um bezüglich eines bestimmten Phänomens repräsentativ zu sein. Um etwa englische Konditionalsätze untersuchen zu können, müsste demnach das Sample 1190 Texte à 2000 Wörter umfassen (also etwa 2,4 Mio. Wörter). Legt man dieses Maß an, erscheinen alle unsere Teilkorpora als ausreichend groß.

[75] Vgl. auch Biber (2010: 242): „The register perspective characterises the typical linguistic features of text varieties, and connects those features functionally to the situational context of the variety."

basieren, was besser zum DeReKo als Datenquelle passt. Die Unterscheidung von Registern ist vor allem im englischsprachigen Raum üblich. Sie wird auch im oben genannten Sinne im LSWE Corpus benutzt. Im deutschsprachigen Raum ist die verwandte, allerdings feinkörnigere Unterscheidung nach Textsorten gängiger. Die Unterteilung des Gesamtkorpus in Register lässt sich nachvollziehen, indem die Teilfaktoren 'Zweck' und 'Entstehung' im Strukturbaum der Abbildung 3 abgearbeitet werden.

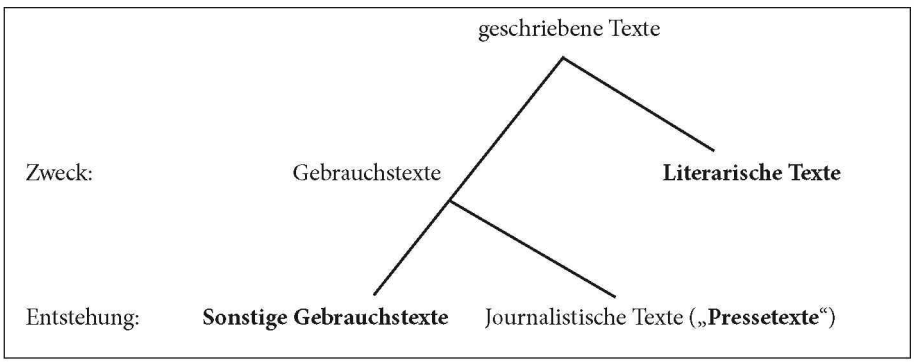



**Abb. 3: Register-Strukturierung des Gesamtkorpus**

Die Zuordnung der Texte zu den einzelnen Registern erfolgte hauptsächlich anhand der DeReKo-Metadaten zu Textsorte/Genre, wie es die Tabelle 9 auf der folgenden Seite zeigt. Allerdings wurden diese an mehreren fraglichen Stellen „von Hand" überprüft, und die Zuordnung wurde gegebenenfalls angepasst.

Als Schwäche dieser Textsorten-/Genre-Klassifizierung kann ausgelegt werden, dass sie manuell durch wechselnde DeReKo-Team-Mitarbeiter vorgenommen wurde und dementsprechend heterogen und stellenweise inkonsistent ausfällt. Andererseits lebt sie gerade – wie die traditionellen Textsortenauffassungen – vom intuitiven Urteil des Sprachbenutzers, ein Aspekt, dem in der Standardsprachlichkeitsproblematik wie in Kapitel 1.1.1 dargestellt eine besondere Rolle zukommt. Um dennoch Idiosynkrasien dieser Einteilung aufzuwiegen, soll ihr in Zukunft eine automatische Stilklassifikation zur Seite gestellt werden.

Das Teilkorpus 'Pressetexte' umfasst ca. 10,5 Mio. Token bzw. ca. 31 000 Texte mehr als das nach 'Medium' unterschiedene Teilkorpus 'Publikumspresse'. Dies liegt daran, dass Beiträge in zahlreichen auf besondere Leserschaft spezialisierten Fachperiodika wie die „Deutsche Richterzeitung" oder die „Naturwissenschaftliche Rundschau" bei der medialen Unterteilung dem Teilkorpus

| Journalistische Texte („Pressetexte“) | Gebrauchstexte | | | Literarische Texte |
|---|---|---|---|---|
| Agenturmeldungen | Abhandlung | Gebetbuch | Prospekt | Aphorismus |
| Anzeigen | Abhandlungen | Gebrauchsanleitung | Protokoll | Autobiografie |
| Diskussionsbeitrag | Abiturarbeit | Geleitwort | Ratgeber | Comic |
| Kolumnen | Anleitung | Gerichtsurteil | Rede | Drama |
| Kommentar | Anmerkung | Gesetz | Resolution | Erzählung |
| Leitartikel | Ansprache | Gesetzesentwurf | Resümee | Erzählungen |
| Leserbrief | Antrag | Gespräch | Rezension | Fabeln |
| Nachrichten | Anweisung | Gnadengesuch | Richtlinie | Filmmanuskript |
| Notiz | Appell | Grundsatzerklärung | Rundfunkbeitrag | Gedicht |
| Presseerklärung | Argumentation | Grußadresse | Rundschreiben | Geschichte |
| Pressemitteilung | Aufforderung | Gutachten | Sachbuch | Hörspiel |
| Vorschlag | Aufklärungstext | Habilitationsschrift | Satzung | Kinderbuch |
| Zeitungsartikel | Aufruf | Handbuch | Schilderung | Kriminalroman |
| | Aufsatz | Handzettel | Schulbuch | Kurzgeschichten |
| Monatsschrift | Ausgabenvermerk | Horoskop | Sprechchöre | Legende |
| Periodikum | Ausschreibung | Information | Statut | Lesebuch |
| Tageszeitung | Bekanntmachung | Informationsblatt | Statuten | Lied |
| Wissenschafts-zeitung | Bemerkung | Informations-material | Stellungnahme | Manifest |
| Wochenzeitung | Beratung | Inhaltsangabe | Studie | Märchen |
| Zeitschrift | Bericht | Inhaltsverzeichnis | Tagebuch | Memoiren |
| Zeitung | Beschluss | Interview | These | Novelle |
| Zeitungsartikel | Beschreibung | Jahrbuch | Thesen | Roman |
| | Bezeichnungs-sammlung | Kochbuch | Titelblatt | Romankapitel |
| | Bibliografie | Konspekt | Titelseite | Sage |
| | Biografie | Konzept | Übersicht | Schauspiel |
| | Botschaft | Lebensdaten | Verfassungs-dokument | Trauerspiel |
| | Brief | Lebenslauf | Verordnung | Unterhaltungs-literatur |
| | Broschüre | Lehrbuch | Vertrag | |
| | Denkschrift | Lexikon | Vorwort | |
| | Diskussion | Lexikonartikel | Verzeichnis | |
| | Dissertation | Losung | Vorschrift | |
| | Dokument | meinungsbildender [Text] | Vorstellung | |
| | Einleitung | Merkblatt | Vortrag | |
| | Empfehlungen | Mitteilung | Wahlwerbung | |
| | Entwurf | Monographie | Werbeprospekt | |
| | Enzyklopädie-Artikel | Nachbemerkung | Werbung | |
| | Erklärung | Nachruf | Wetterbericht | |
| | Erlass | Nachwort | Wörterbuch | |
| | Erwiderung | Packungsbeilage | Wörterbuchartikel | |
| | Essay | Patent | Worterklärung | |
| | Exzerpt | Personenregister | | |
| | Fernsehsendung | Petition | | |
| | Flugblatt | Postulat | | |
| | Flugschrift | Predigt | | |
| | Forderung | Problemkatalog | | |
| | Formular | Produktbeschrei-bung | | |
| | Forschungsbericht | Programm | | |
| | Fußnote | | | |

**Tab. 9: Zuordnung der Textsorten und Genres der DeReKo-Metadatenbank zu Registern**

'Sonstige Printmedien' zugeordnet wurden, um sie von Texten aus Tageszeitungen, populären wöchentlichen Nachrichtenmagazinen und Publikumszeitschriften unterscheiden zu können.

### Staat und Region

| Staat | Token | Anteil | Texte | Anteil |
|---|---|---|---|---|
| Deutschland (D) | 3 145 604 245 | 72,84% | 11 093 116 | 69,61% |
| Deutschland – Ost (D-O) | 2 786 747 | 0,06% | 1 707 | 0,01% |
| Deutschland – West (D-W) | 35 536 911 | 0,82% | 73 471 | 0,46% |
| Schweiz (CH) | 376 112 826 | 8,71% | 1 204 280 | 7,56% |
| Österreich (A) | 733 690 278 | 16,99% | 3 489 180 | 21,89% |
| unklassifizierbar | 25 056 667 | 0,58% | 75 326 | 0,47% |
| Gesamt | 4 318 787 674 | 100,00% | 15 937 080 | 100,00% |

Tab. 10: Unterteilung des Gesamtkorpus im Hinblick auf 'Staat'

Die räumliche Strukturierung des Untersuchungskorpus beruht auf zwei Parametern. Bei **Staat** konnten teilweise direkt die DeReKo-Metadaten genutzt werden, wobei punktuell Korrekturen vorgenommen werden mussten (z.B. die DeReKo-Zuordnung „D" bei in Deutschland herausgegeben Werken österreichischer und schweizerischer Autoren wurde in „A" bzw. „CH" geändert). Darüber hinaus wurden für die Zeit bis zur und einschließlich der Wende[76] die Kategorien 'Deutschland – West' („D-W") und 'Deutschland – Ost' („D-O") eingerichtet. Texte aus Deutschland, für die keine genauere Zuordnung möglich war bzw. die in die Zeit nach der Wende fallen, wurden der Kategorie 'Deutschland' („D") zugewiesen.

Die Unterteilung im Hinblick auf **Region** richtet sich nach traditionellen dialektgeografisch fundierten Auffassungen wie sie z.B. König (2004: 230f.) und Wiesinger (1983: 807ff.) zugrunde liegen.[77] Bei Zuordnungsproblemen erfolgte der Abgleich mit der Karte 47.4 von Wiesinger (1983).

[76] Mit „Wende" ist die Zeit von 1989 bis 1990 gemeint, in die alle Dokumente fallen, die in den im IDS manuell zusammengestellten Wendekorpora (wk) enthalten sind.

[77] Die Regionen in Tabelle 11 entsprechen dabei den üblichen Sprachlandschaftsbezeichnungen in folgender Reihenfolge: Westniederdeutsch, Ostniederdeutsch, Westmitteldeutsch, Ostmitteldeutsch, Nordoberdeutsch (Ostfränkisch), Westoberdeutsch und Ostoberdeutsch.

| Region | Token | Anteil | Texte | Anteil |
|---|---|---|---|---|
| Nordwest | 185 211 687 | 4,29% | 953 125 | 5,98% |
| Nordost | 266 000 795 | 6,16% | 1 086 675 | 6,82% |
| Mittelwest | 1 095 986 696 | 25,38% | 4 334 307 | 27,20% |
| Mittelost | 1 563 144 | 0,04% | 562 | 0,004% |
| Mittelsüd | 279 140 991 | 6,46% | 860 829 | 5,40% |
| Südwest (einschl. Schweiz) | 377 135 281 | 8,73% | 1 288 257 | 8,08% |
| Südost (einschl. Österreich) | 653 769 099 | 15,14% | 3 003 943 | 18,85% |
| überregional | 1 117 866 209 | 25,88% | 3 097 054 | 19,43% |
| unklassifizierbar | 342 113 772 | 7,92% | 1 312 328 | 8,23% |
| Gesamt | 4 318 787 674 | 100,00% | 15 937 080 | 100,00% |

**Tab. 11: Unterteilung des Gesamtkorpus im Hinblick auf 'Region'**

Es wurde versucht, alle Texte nach beiden Parametern zu klassifizieren. Nach dem Parameter 'Staat' konnten fast alle Texte klassifiziert werden, nach dem Parameter 'Region' aber nur ca. 92%. Die wichtigste Grundlage für die räumliche Klassifizierung Angaben zu Titel, Autor und Erscheinungsort. Bei Texten, deren Autoren der Texte bildeten die in den Metadaten verzeichneten entweder in den Metadaten angegeben oder anderweitig zu ermitteln waren, richtete sich die Zuordnung nach dem Wirkungsort bzw. dem Heimatort des Autors (eventuelle Unterschiede zwischen den beiden wurden in den Zusatzinformationen zum Text festgehalten). Beim Parameter 'Staat', der politgeografisch zu verstehen ist, war – insbesondere bei der Unterscheidung zwischen 'D-W' und 'D-O' – der Wirkungsort des Autors entscheidend, beim Parameter 'Region' dagegen die sprachliche Herkunft des Autors. Daraus ergibt sich für die Texte bis zur Wende unter anderem, dass die Werke von Siegfried Lenz die 'Staat'-Kategorie 'D-W' und die Werke von Christa Wolf die 'Staat'-Kategorie 'D-O' zugewiesen bekamen, im Hinblick auf 'Region' aber den Texten beider Autoren aufgrund deren sprachlicher Herkunft die Kategorie 'Nordost' zugeordnet wurde.

Bei Pressetexten (deren Autoren in den allermeisten Fällen nicht zu ermitteln waren) wurden in der Regel ganze Zeitungen bzw. Zeitschriften als zu einer bestimmten Region zugehörig oder als überregional klassifiziert und einem Staat zugewiesen:

**Überregionale Presse, Deutschland**

*Computer Zeitung*
*Der Spiegel*
*die tageszeitung*
*Die Zeit*
*Frankfurter Allgemeine Zeitung*
*Frankfurter Rundschau*
*Magazin Lufthansa Bordbuch / Deutsch*
*Meldungen der Deutsche Presse-Agentur*
*spektrumdirekt*
*Süddeutsche Zeitung*
*VDI Nachrichten*

**Überregionale Presse, Deutschland – West**

*ADAC Motorwelt*
*Bild am Sonntag*
*Bild der Wissenschaft*
*Bildzeitung*
*Das ÖTV-Magazin*
*Das Parlament*
*Deutsche Apothekerzeitung*
*Deutsche Handwerkszeitung*
*Deutsche Richterzeitung*
*Deutsche Volkszeitung*
*Deutsches Allgemeines Sonntagsblatt*
*Die Belegschaft*
*Die Bundeswehr*
*Die Welt*
*Europäische Grundrechte Zeitschrift*
*ExtraBlatt*
*GEO*
*Handelsblatt*
*Kirche und Kommune*

*Kosmos – Das Naturmagazin*
*Libertas. Europäische Zeitschrift*
*Mieterzeitung*
*Monatsschrift Kinderheilkunde*
*Natur*
*Naturwissenschaftliche Rundschau*
*neue deutsche schule*
*stern*
*unsere zeit*
*Vereinigte Linke*
*Wirtschaftswoche*
*Zeitschrift für medizinische Laboratoriumsdiagnostik*
*Zeitschrift für Allgemeinmedizin*
*Zeitschrift für Hautkrankheiten*
*Zeitschrift für klinische Medizin*
*Zeitschrift für Physiotherapie*
*Zeitschrift für Rechtspolitik*

sowie Jahrgänge der *FAZ*, des *Spiegel* und der *taz* vor 1991

**Überregionale Presse, Deutschland – Ost**

*Der Morgen*
*Der Morgen zur Wahl*
*Junge Welt*
*Neue Zeit*
*Neues Deutschland*
*stern extra zur Wahl*
*telegraph*
*Wochenpost*

**Regionale Presse**

Nordwest, Deutschland:

*Braunschweiger Zeitung* (Braunschweig)
*Die Norddeutsche* (*Weser Zeitung*) (Bremen)
*Hamburger Morgenpost* (Hamburg)
*Hannoversche Allgemeine* (Hannover)
*Hildesheimer Allgemeine Zeitung* (Hildesheim)
*Kieler Nachrichten* (Kiel)

*Neue Osnabrücker Zeitung* (Osnabrück)
*St. Pauli Nachrichten* (Hamburg)

Nordost, Deutschland:

*Berliner Morgenpost* (Berlin)
*Berliner Zeitung* (Berlin)
*Berliner Rundschau* (Berlin)
*Berliner StattBlatt* (Berlin)
*Der Tagesspiegel* (Berlin)
*Pankower Spiegel* (Berlin)

Mittelwest, Deutschland:

*Die Rheinpfalz* (Ludwigshafen)
*Kölner Stadtanzeiger* (Köln)
*Mannheimer Morgen* (Mannheim)
*Rhein-Zeitung* (Koblenz)
*Rheinischer Merkur* (Bonn)
*Saarbrücker Zeitung* (Saarbrücken)

Mittelost, Deutschland:

*DAZ = Die Leipziger andere Zeitung* (Leipzig)
*Die Leipziger andere Zeitung = DAZ*
*Die Union* (Dresden)
*DSU-Kurier* (Leipzig)
*Leipziger Volkszeitung* (Leipzig)
*Sachsen-Spiegel* (Dresden)
*SPD. Dresden aktuell* (Dresden)

Mittelsüd, Deutschland:

*Frankenpost* (Hof)
*Nürnberger Nachrichten* (Nürnberg)
*Nürnberger Zeitung* (Nürnberg)

Südwest, Deutschland:

*Stuttgarter Zeitung* (Stuttgart)
*Zollern-Alb-Kurier* (Balingen)

Südwest, Schweiz:

*Die Südostschweiz* (Chur)
*Neue Zürcher Zeitung* (Zürich)
*St. Galler Tagblatt* (St. Gallen)
*Schweizerische Juristen-Zeitung* (Zürich)
*Swiss Biotech* (Zürich)
*Zürcher Tagesanzeiger* (Zürich)

Südwest, Österreich:

*Vorarlberger Nachrichten* (Schwarzach)

Südost, Deutschland:

*Bayernkurier* (München)
*Freisinger Tagblatt* (Freising)

Südost, Österreich:

*Burgenländische Volkszeitung* (Burgenland)
*Die Presse* (Wien)
*Kleine Zeitung* (Graz)
(*Neue*) *Kronen-Zeitung* (Wien)
*Niederösterreichische Nachrichten* (Sankt Pölten)
*Oberösterreichische Nachrichten* (Linz)
*Salzburger Nachrichten* (Salzburg)
*Tiroler Tageszeitung* (Innsbruck)
*Wiener klinische Wochenschrift* (Wien)

Hinsichtlich der pauschalen räumlichen Klassifizierung von Zeitungen und Zeitschriften könnte hier eingewendet werden, dass die Autoren ein und desselben Presseorgans ganz unterschiedlicher sprachlicher Herkunft sein können. Es lässt sich allerdings wohl trotzdem davon ausgehen, dass etwa in einer Schweizer Zeitung insgesamt der südwestliche „Einschlag" überwiegt und auch ein Berliner Autor, wenn er für diese schreibt, sich anzupassen versucht oder zumindest mit Berlinismen zurückhält.

**Domäne**

| Domäne | Token | Anteil | Texte | Anteil |
|---|---|---|---|---|
| Fiktion | 9 183 868 | 0,21% | 4 182 | 0,03% |
| Kultur/Unterhaltung | 1 786 027 342 | 41,35% | 6 645 823 | 41,70% |
| Mensch/Natur | 96 314 485 | 2,23% | 410 549 | 2,58% |
| Politik/Wirtschaft/Gesellschaft | 1 850 089 257 | 42,84% | 6 547 047 | 41,08% |
| Technik/Wissenschaft | 297 756 113 | 6,89% | 1 303 901 | 8,18% |
| unklassifizierbar | 279 421 366 | 6,47% | 1 025 582 | 6,47% |
| Gesamt | 4 318 787 674 | 100,00% | 15 937 080 | 100,00% |

Tab. 12: Unterteilung des Gesamtkorpus im Hinblick auf 'Domäne'

Eine Gliederung des Korpus nach inhaltlichen Domänen ist schon aus dem BNC XML bekannt (vgl. Kapitel 1.3). Trotz der Einwände gegen inhaltliche Klassifikationen, wie sie Sinclair (2004) äußert, liegt eine solche Klassifikation für das vorliegende Projekt auf der Hand, da für das DeReKo bereits fast „flächendeckend" Informationen zur inhaltlichen Ausrichtung einzelner Texte vorliegen. Diese Informationen wurden in einem automatischen Klassifikationsverfahren, das im Programmbereich Korpuslinguistik des IDS erarbeitet wurde (Genaueres unter http://www.ids-mannheim.de/kl/projekte/methoden/te.html und in Weiß 2005) den Texten hinzugefügt. In den einzelnen 'Domänen' werden hier die dort semiautomatisch ermittelten Themen zusammengefasst. Unter 'Fiktion' werden in unserem Vorhaben zusätzlich alle Texte versammelt, die dem Register 'Literarisches' angehören. Sie werden nicht thematisch klassifiziert, weil eine solche Klassifikation im Falle von fiktionalen bzw. „kreativen" Texten als wenig sinnvoll erscheint (ähnlich BNC User Reference Guide).[78] Wie sich die übrigen Domänen aus DeReKo-Themen zusammensetzen, zeigt Tabelle 13.[79]

[78] http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html (Stand: 2.7.2013).

[79] Zur feinkörnigen Zusammensetzung der Themenkategorien im DeReKo vgl. http://www1.ids-mannheim.de/kl/projekte/methoden/te/themen_short.pdf (Stand: 2.7.2013).

| Domäne | Themenkategorien im DeReKo |
|---|---|
| Kultur/Unterhaltung | Freizeit/Unterhaltung, Kultur, Sport |
| Mensch/Natur | Gesundheit/Ernährung, Natur/Umwelt |
| Politik/Wirtschaft/Gesellschaft | Politik, Staat/Gesellschaft, Wirtschaft/Finanzen |
| Technik/Wissenschaft | Technik/Industrie, Wissenschaft |

**Tab. 13: Zusammensetzung der nicht fiktionalen Domänen**

### 1.4.4 Ausgewogenes Korpus

Das ausgewogene Korpus soll aus ausgewählten Texten des Gesamtkorpus zusammengestellt und ausgewogen in Bezug auf den Parameter 'Medium' sein. Die Konstellation der Untersuchungskorpora des Projekts erinnert also an das BNC Baby mit dem 100 Mio. Wörter großen BNC XML im Hintergrund. Das BNC Baby setzt sich aus vier Teilkorpora zusammen, die jeweils ein Genre repräsentieren und eine Million Wörter umfassen. Das ausgewogene Korpus unseres Vorhabens umfasst fünf Teilkorpora, für die die Größe von jeweils 5 Mio. Wörtern anvisiert wird. Damit soll sichergestellt werden, dass die Teilkorpora groß genug sind, um aussagekräftige Analysen zur Grammatik im jeweiligen Bereich zu erlauben. Als Orientierung dienen dabei die Überlegungen von Biber (1993) zu Mindestgrößen von Korpora für Analysen von grammatischen Erscheinungen sowie das LSWE Corpus, in dem sich die Teilkorpora zu einzelnen Registern in der Größe von jeweils ca. 5 Mio. Wörtern (Biber et al. 1999, 2006: 24) bei Untersuchungen zur Grammatik des Standardenglischen bewährt haben. Die Zielgröße der Teilkorpora ist in der momentanen Ausbaustufe des ausgewogenen Korpus noch nicht erreicht (vgl. Tabelle 14), denn zurzeit sind passende Texte aus dem Bereich 'Gesprochenes' im DeReKo noch nicht in ausreichender Anzahl vorhanden. Schon jetzt weisen die Teilkorpora aber eine willkommene breite Streuung im Hinblick auf Zeit sowie Regionen, Textsorten und Domänen auf.

| Medium | Token | Anteil | Texte | Anteil |
|---|---|---|---|---|
| Publikumspresse | 4025846 | 20,10% | 8537 | 42,37% |
| Bücher | 4000385 | 19,98% | 350 | 1,74% |
| Sonstige Printmedien | 4000020 | 19,97% | 1764 | 8,76% |
| Internet/Wikipedia | 4000102 | 19,98% | 6897 | 12,90% |
| Gesprochenes | 4000404 | 19,97% | 2600 | 34,23% |
| Gesamt | 20026757 | 100,00% | 20148 | 100,00% |

**Tab. 14: Zusammensetzung des ausgewogenen Korpus**

Die Teilkorpora umfassen zurzeit jeweils ca. 4 Mio. Token. Für das Teilkorpus 'Bücher' werden Texte manuell aus DeReKo ausgewählt, wobei die Metadaten noch einmal sorgfältig auf ihre Richtigkeit und Vollständigkeit überprüft werden. Innerhalb des Teilkorpus wird Ausgewogenheit in Bezug auf den Parameter 'Register' angestrebt, sodass literarische Texte und Gebrauchstexte aktuell jeweils ca. 2 Mio. Wörter umfassen. Das Teilkorpus weist schon jetzt eine gute Streuung in Bezug auf die Parameter Zeit, Region, Register und Domäne auf.

Bei der Zusammenstellung des Teilkorpus 'Publikumspresse' wird eine Ausgewogenheit im Hinblick auf den Parameter 'Region' sowie im Hinblick auf die Erscheinungsdekade (1990er-Jahre versus 2000er-Jahre) angestrebt (vgl. Tabelle 15). Bei jeder Region zuzüglich der Gruppe 'Überregional' werden 250000 Wörter pro Dekade angestrebt, die mittels geschichteter Stichprobe zufällig ausgewählten Texten entstammen. Diese Vorgabe kann nur für die Regionen 'Mittelost' und 'Nordwest' nicht vollständig erfüllt werden, sodass zusätzliche überregionale Texte hinzugezogen werden müssen, um die vorläufige Gesamtgröße des Teilkorpus von 4. Mio. Wörtern zu erreichen.

| Region | Wörter pro Dekade | | Wörter gesamt |
|---|---|---|---|
| | 1990-1999 | 2000-2009 | |
| Mittelost | 96116 | | **96116** |
| Mittelsüd | 250019 | 250537 | **500556** |
| Mittelwest | 251400 | 250552 | **501952** |
| Nordost | 176281 | 250215 | **426496** |
| Nordwest | | 250799 | **250799** |
| Südost | 252285 | 251713 | **503998** |
| Südwest | 250542 | 202301 | **452843** |
| überregional | 651988 | 641098 | **1293086** |
| **Gesamt** | **1928631** | **2097215** | **4025846** |

**Tab. 15: Zusammensetzung des Teilkorpus 'Publikumspresse'**

Beim Teilkorpus **Internet/Wikipedia** schließlich erfolgt eine Zufallsauswahl der Texte aus der Ausgabe 2005 von Wikipedia.

Somit eignet sich das ausgewogene Korpus in besonderer Weise für Untersuchungen im Hinblick auf den Parameter 'Medium'. Es verspricht aber auch eine zuverlässige Erfassung von Register- sowie zeitlichen und regionalen Unterschieden und macht schließlich eine unbeabsichtigte Übergewichtung von einzelnen Textsorten und Domänen unwahrscheinlich.

## 1.5 Zur Erinnerung

Es wurde schon vielerorts[80] betont, soll aber dennoch noch einmal gesagt werden: Feststellungen zum Verhalten von Phänomenen in einem Korpus bleiben immer in erster Linie Feststellungen über dieses konkrete Korpus. Bei jeder Verallgemeinerung der Ergebnisse muss in Erwägung gezogen werden, dass sich die Sachlage ungeachtet avancierter mathematischer Auswertungsmethoden (vgl. Kapitel 3) bei jeder anderen Korpuszusammenstellung im Prinzip auch anders gestalten könnte.

[80] Z.B. Hunston (2002: 23), Clancy (2010: 87).

# 2. Die Korpusdatenbank *KoGra-DB*

Der nachfolgend vorgestellte Ansatz zu Design und Implementierung eines korpusgrammatischen Verwaltungs- und Abfragesystems nimmt sich verschiedener im Kontext der maschinellen Erschließung sehr großer Sprachkorpora („very large corpora") aktuellen wissenschaftspraktischen Problematiken an. Aus informationstechnologischer Perspektive wird ein neuartiges funktionales Verfahren für die Korpusrecherche unter Einbeziehung komplexer Mehrebenen-Annotationen sowie text- und korpusspezifischer Metadaten vorgestellt.

## 2.1 Grundlegende technische Weichenstellungen

In der Einleitung zu diesem Band wurde bereits aufgezeigt, dass und warum sich die korpusbasierte quantitative Untersuchung natürlichsprachlicher Phänomene zunehmend als anerkanntes linguistisches Paradigma etabliert. Nicht nur in der Grammatikforschung, sondern gleichermaßen in der Pragmatik, der Semantik oder der Lexikologie dienen Sprachkorpora als empirische Basis für die Formulierung und Überprüfung von Hypothesen oder für die Entwicklung computerlinguistischer Anwendungen. Um diesem Anspruch gerecht zu werden, bestehen linguistisch motivierte Korpora nicht allein aus den zumeist textuellen – bzw. in Abhängigkeit von Untersuchungsgegenstand und Erkenntnisinteresse gegebenenfalls auch gesprochenen – Primärdaten, sondern ganz wesentlich zusätzlich aus (morpho-)syntaktischen, semantischen, phonetischen oder sonstigen Informationen über diese Daten; entsprechende Anreicherungen werden gemeinhin in Form von Mehrebenen-Annotationen kodiert. Angesichts der Fülle digital erschlossener Primärdaten werden Annotationen zunehmend mit Hilfe maschineller Parser und Tagger generiert, wobei für die vergleichende Auswertung und zum Erkennen problematischer Etikettierungen oft der parallele Einsatz mehrerer dieser Werkzeuge angeraten erscheint. Hinzu kommen korpus- oder textspezifische außersprachliche Klassifizierungen wie Medium, Domäne, Region, Jahr etc., die insbesondere bei vielschichtigen Vorhaben wie dem hier vorgestellten zur grammatischen Variation im Standarddeutschen unverzichtbar sind.

Zusammenfassend kann konstatiert werden: Die für linguistische Studien verfügbare und mithin zu analysierende Datenmenge nimmt seit Jahren kontinuierlich zu.

Diese aus empirischer Sicht („more data are better data")[1] begrüßenswerte Entwicklung hat ihren Preis: Korpusrecherchesysteme – das unverzichtbare „Handwerkszeug" der Korpuslinguisten – entwickeln sich nolens volens zu spezialisierten Produkten, die sich von traditionellen Volltextsuchen und Information Retrieval (IR)-Systemen dadurch unterscheiden, dass sie nicht allein die horizontale Verkettung von Sprachelementen abdecken, sondern sämtliche verfügbaren Annotationen der unterschiedlichen linguistischen Beschreibungsebenen sowie die außersprachlichen Metadaten kombiniert erschließbar machen sollen. Eine Reihe einschlägiger Forschungsarbeiten beschäftigen sich infolgedessen mit der Repräsentation von Mehrebenen-Annotationen sowie mit der Mächtigkeit und Adäquatheit korpuslinguistischer Abfragesprachen.[2] Die Optimierung der Datenhaltung und – darauf aufbauend – die Implementierung performanter ebenenübergreifender Suchanfragen mit mehreren Suchprädikaten oder gar regulären Ausdrücken erscheinen derzeit dagegen als eher spärlich erforscht; Kapitel 2.3 geht auf einige existierende Arbeiten ein. Doch an genau diesem Punkt stoßen etablierte Systeme an ihre Grenzen.

Vereinfacht gesagt: Die durch die rapide hard- und softwaretechnische Weiterentwicklung von Speichersystemen und Tagging-Werkzeugen verfügbare Menge an linguistisch relevanten Daten übersteigt mittlerweile oft unsere Möglichkeiten zur gezielten zeitnahen Auswertung.

Exemplarisch für das Wachstum aktueller Forschungskorpora sei der Ausbau des Deutschen Referenzkorpus DeReKo erwähnt, der primären Datenquelle für den Aufbau der Projektkorpora: Was in den Sechziger- und Siebzigerjahren des vorigen Jahrhunderts als vergleichsweise bescheidene Sammlung elektronischer Texte – anfangs noch auf Lochkarten – am Institut für Deutsche Sprache (IDS) begann, hat sich im Verlauf der vergangenen Jahrzehnte zur weltweit umfangreichsten digitalen Sammlung deutschsprachiger Texte für sprachwissenschaftliche Untersuchungen entwickelt. Der Umfang von DeReKo erhöhte sich seit 1992 von ca. 28 Millionen auf über 5 Milliarden Textwörter im Jahre 2012 (vgl. Abbildung 1), das entspricht über 12 Millionen Buchseiten, wenn man durchschnittlich 400 Wörter pro Seite zugrunde legt.[3] Sämtliche Texte sind in einem einheitlichen Strukturformat (IDS-XCES) kodiert, mit textspezifischen Merkmalen angereichert und parallel mit drei

[1] Vgl. Church/Mercer (1993).

[2] Vgl. z.B. Rehm et al. (2008), Zeldes et al. (2009), Kepser et al. (2010).

[3] Diese Angaben sowie die Grafik wurden folgender Internetseite zum DeReKo-Korpusausbau entnommen: http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html (Stand: August 2013).

Tagging-Werkzeugen morphosyntaktisch annotiert (TreeTagger,[4] Connexor Machinese Phrase Tagger,[5] Xerox Incremental Parser[6]). Gemeinsam nehmen diese Daten ca. 3 Terabyte (TB) Speicherplatz ein, wobei sich das Datenvolumen durch die kontinuierliche Akquise zusätzlicher Teilkorpora auch zukünftig weiter erhöhen wird. Im Rahmen der Korpusgrammatik-Projektarbeit sind darüber hinaus sowohl das punktuelle Hinzufügen bestimmter Mediengattungen (z.B. Online-Inhalte oder verschriftete Sprechsprache) als auch die Ergänzung um zusätzliche (beispielsweise syntaktische) Annotationsebenen geplant; vgl. hierzu auch die Beschreibung der Korpusbasis in Kapitel 1.4.

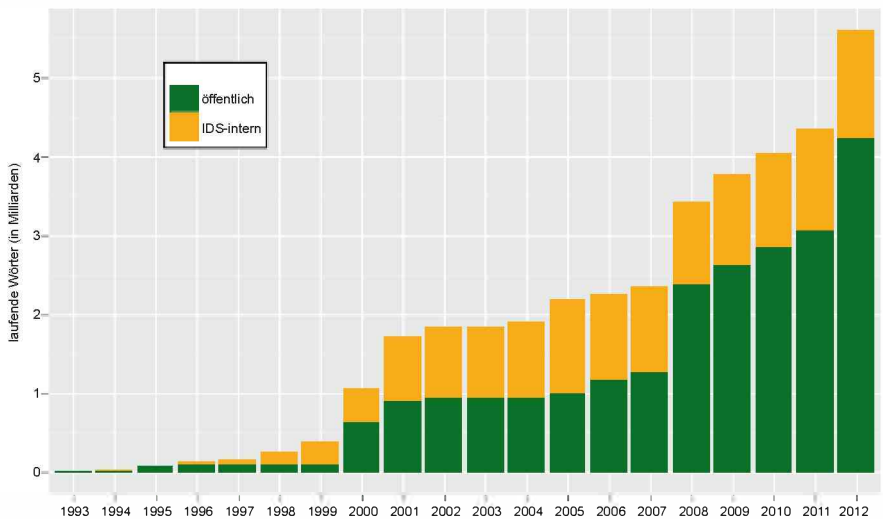



**Abb. 1: Entwicklung des DeReKo-Gesamtumfangs**

Um eine solche Menge an Primär-, Annotations- und Metadaten hinsichtlich der im Projektkontext anstehenden linguistisch-grammatischen Fragestellungen angemessen und reproduzierbar auswerten zu können, bedarf es eines gleichermaßen verlässlichen, performanten und funktional erweiterbaren Ansatzes. Dieser sollte außerdem aus Gründen der Nachhaltigkeit grundsätzlich kompatibel mit Standards sein, die gegenwärtig im Kontext nationaler und supranationaler Infrastrukturverbundprojekte (z.B. CLARIN, DARIAH, TextGrid)[7] ausgearbeitet werden. Im computerlinguistischen bzw. informations-

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (Stand: August 2013).

[5] http://www.connexor.eu/technology/machinese/index.html (Stand: August 2013).

[6] http://open.xerox.com/Pages/Linguistic%20Technologies (Stand: August 2013).

[7] http://www.clarin.eu, http://www.dariah.eu und http://www.textgrid.de.

technologischen Forschungsumfeld finden sich hierzu mehrere einschlägige, technologisch teilweise überlappende und kombinierbare Modelle:

a) **Dateibasierte Lösungen:** Primärdaten sowie die Ergebnisse linguistischer Tagger liegen gemeinhin in Form strukturierter XML-Dateien vor. Ein naheliegender Ansatz ist deshalb, diese Daten mit existierenden XML-Werkzeugen und unter Ausnutzung etablierter Standards wie XPath (XML Path Language zur Adressierung von Teilen eines XML-Dokuments), XSLT (Extensible Stylesheet Language Transformations zur Konvertierung zwischen unterschiedlichen XML-Sprachen) und XQuery (XML Query Language zur Recherche in XML-Dokumenten) zu analysieren. Allerdings sind dieser Vorgehensweise vergleichweise enge Grenzen gesetzt, und zwar gleichermaßen hinsichtlich der Datenmenge – sämtliche von uns getesteten Standalone-XML-Parser und -Prozessoren (SAX, Xerces, Microsoft XML Parser) kapitulierten vor dem umfangreichen Output der linguistischen Tagger – als auch in Bezug auf die performante parallele Abfrage verschiedenartiger XML-Strukturen, wie sie bei Mehrebenen-Annotationen üblich sind.

b) **In-Memory-Modelle:** Hier reicht die Palette von proprietären Softwareentwicklungen bis hin zu kommerziellen In-Memory-Datenbanken (IMDB). Der gemeinsame Lösungsansatz beruht auf der Speicherung sämtlicher relevanten Daten im schnellen Hauptspeicher (RAM) eines Rechners sowie auf der damit verbundenen drastischen Reduzierung der Retrievalzeiten, unabhängig vom letztlich eingesetzten Query-Algorithmus. Ad-hoc-Abfagen kleiner und mittlerer Datenmengen sind auf diese Weise performant durchführbar, während die Verwaltung sehr großer Sprachkorpora im Terabyte-Bereich – selbst beim Einsatz von Kompressionstechniken – derzeit kaum machbar erscheint.

c) **„Klassische" IR-Systeme:** Spezialisierte Systeme für das Information Retrieval wie beispielsweise Apache Lucene, Oracle Text oder Managing Gigabytes[8] unterstützen die flexible Suche nach komplexen Textinhalten. Dokumente werden intern in eine systemspezifische Repräsentationsform umgewandelt, indiziert und unter Zuhilfenahme von Abfragemodellen wie dem Boole'schen oder dem Vektorraum-Modell recherchierbar gemacht.

[8] Apache Lucene ist eine gleichermaßen populäre und leistungsstarke freie Software zur Volltextsuche, die aufgrund ihrer Skalierbarkeit in vielen großen Internetportalen (z.B. Twitter) zum Einsatz kommt; vgl. http://lucene.apache.org. Oracle Text bietet vielfältige Index- und Suchanfragetypen für die Verwaltung semistrukturierter Texte in relationalen Datenbanken; vgl. http://www.oracle.com/technetwork/. Auf Managing Gigabytes basiert z.B. die Verwaltung und Abfrage von DeReKo am Institut für Deutsche Sprache (IDS) vermittels COSMAS II; vgl. http://www.ids-mannheim.de/cosmas2/.

Auch semi-strukturierte XML-Dokumente lassen sich auf diese Weise unter Einbeziehung einzelner Auszeichnungshierarchien durchsuchen. Der originäre Anwendungsfokus liegt primär auf der horizontalen Verkettung von Sprachelementen und weniger auf der (vertikalen) Recherche in Mehrebenen-Annotationen.

d) **DBMS-basierte Konzepte:** Datenbankmanagementsysteme (DBMS) arbeiten als zusätzliche Softwareschicht zwischen Betriebssystem und Anwendung und bieten neben einer grundlegenden Unterstützung umfangreicher Datensammlungen (Stichworte: Datensicherheit, Datenintegrität, Mehrbenutzerfähigkeit, verteilte und physikalisch praktisch unlimitierte Datenhaltung, integrierte Caching-Mechanismen und Optimierer) insbesondere präzise und wohldefinierte Abfragesprachen wie die Structured Query Language (SQL). Weiter verbreitet ist das relationale Datenmodell, das in den meisten relevanten Implementierungen um objektorientierte oder XML-Funktionalitäten erweitert wurde; daneben rücken in jüngster Zeit sogenannte NoSQL („Not only SQL")-Systeme in den Fokus strukturierter Datenspeicherung.

Mehrere aktuelle sprachwissenschaftliche Korpusprojekte – in Deutschland beispielsweise das Wortschatz-Projekt[9] oder ANNIS[10] – setzen softwareseitig bereits auf relationale DBMS. In Anbetracht der sehr großen zu analysierenden Datenmenge sowie der Komplexität der zu berücksichtigen Mehrebenen-Annotationen liegt analog hierzu auch für die Erforschung grammatischer Variationen die Verwendung eines (objekt-relationalen und XML-fähigen) DBMS nahe. Unser darauf basierendes Verwaltungs- und Abfragesystem wird im Folgenden Korpusgrammatik-Datenbank (*KoGra-DB*) genannt.

Anknüpfend an die Entscheidung pro RDBMS stellt sich unmittelbar die Frage, welcher Art die interne Repräsentation der relevanten Primär-, Annotations- und Metadaten sein sollte. Zwar liegen diese größtenteils bereits in Form strukturierter XML-Dateien vor, unsere Erfahrungen mit deren Weiterverarbeitung begründen jedoch die Annahme, dass XML im Kontext der Korpusannotation eher als internes bzw. externes Austauschformat denn als Grundlage für ein effizientes Retrieval dienen sollte. Es gilt also, die Quelldaten im Zuge des Imports in die *KoGra-DB* so zu relationieren, dass in einem nächsten Schritt präzise, zielführende und performante SQL-Abfragen durchführbar sind.

[9] Vgl. http://wortschatz.uni-leipzig.de.

[10] Vgl. http://www.sfb632.uni-potsdam.de/d1/annis/ (Stand: August 2013).

## 2.2 Quelldaten: Import und Aufbereitung

Die aktuellen Inhalte der *KoGra-DB* entstammen der DeReKo-Freigabe vom 20.03.2011[11] und liegen vor dem Import als wohlgeformte XML-Dateien vor. Jeweils eine XML-Datei fasst die Inhalte eines DeReKo-Einzelkorpus – bei Zeitungen/Zeitschriften o.Ä. für jeweils ein Kalenderjahr – zusammen, wobei ein Korpus stets aus einem oder mehreren Dokumenten und jedes Dokument aus einem oder mehreren Einzeltexten bestehen kann. Jeder Text ist ist durch eine eindeutige Sigle (Buchstaben-Ziffern-Kombination, z.B. A97/APR.00001) charakterisiert, die wiederum ein Kürzel des betreffenden Korpus enthält. Metadaten für die Pflege der DeReKo-Korpora verwaltet der Programmbereich Korpuslinguistik am IDS in einer separaten Datenbank.[12]

### 2.2.1 Textspezifische Metadaten

Textspezifische Metadaten importieren wir aus zwei Quellen, nämlich aus den DeReKo-Primärtexten im IDS-XCES-Format[13] und aus der DeReKo-Metadatenbank. Die Ergebnisse werden aggregiert, ausgewertet und anschließend unter Zuhilfenahme maßgeschneiderter Skripts teilautomatisch angepasst bzw. ergänzt. Aus der DeReKo-Metadatenbank werden folgende textspezifische Angaben übernommen:

a) Textsorte, Genre
   Ausprägungen: Abhandlung, Agenturmeldung, Beratung, Bericht, Zeitung etc.

b) Ressort
   Ausprägungen: Astronomie, Auto, Regional, Sport, Unterhaltung etc.

c) Topic
   Ausprägungen: Freizeit, Unterhaltung, Reisen, Fiktion, Vermischtes etc.[14]

d) Staat
   Ausprägungen: A, CH, D sowie nachgeführt Deutschland – Ost (D-O), Deutschland – West (D-W), vgl. Kapitel 1.4.

[11] Institut für Deutsche Sprache (2011); zu Korpusstruktur und IDS-Textmodell siehe http://www.ids-mannheim.de/kl/projekte/korpora/textmodell.html (Stand: August 2013).

[12] Kupietz/Keibel (2009) bieten eine Übersicht der besagten DeReKo-Metadaten.

[13] Zum IDS-Textmodell und dem auf XCES (Corpus Encoding Standard for XML) basierenden IDS-XCES siehe http://www.ids-mannheim.de/kl/projekte/korpora/idsxces.html (Stand: August 2013).

[14] Weiß (2005) beschreibt die thematische DeReKo-Erschließung in einem semiautomatischen Verfahren, welches die Anwendung von Textmining (Dokumentclustering) und die manuelle Zuordnung von Clustern in eine externe Ontologie beinhaltet.

Die XCES-Primärtexte enthalten – neben Auszeichnungen zu Textstruktur und -layout, die für unsere Zielsetzungen nicht relevant erscheinen – ebenfalls eine Reihe auswertungswürdiger Angaben (siehe Listing 1), von denen wir folgende extrahieren:

e) Titel

f) Autorenname (als Hilfsmittel z.B. für regionale Zuordnungen)

g) Texttyp
   Ausprägungen: Bericht, Interview, Kommentar, Tageszeitung, Zitat etc.

h) Publikationsdatum
   Ausprägungen: Jahreszahl bzw. Jahrzehnt; bei literarischen Werken ist oft anstelle des Ersterscheinungsdatums das Publikationsdatum der spezifischen Auflage enthalten. Wir fügen deshalb gegebenenfalls das Ersterscheinungsdatum hinzu.

i) Publikationsort

Bedingt durch die im IDS-Textmodell definierte Trennung zwischen Text-, Dokument- und Korpusebene können deskriptive Metadaten an unterschiedlichen Stellen in der XML-Hierarchie erscheinen, was Konsequenzen für die Algorithmisierung des Imports in die *KoGra-DB* hat: Findet die Importprozedur beispielsweise in einem der Textsigle folgenden Autorenelement keinen Inhalt, so überprüft sie sukzessive zunächst den zum übergeordneten Dokument gehörigen Metadatenblock und, sofern auch dort nichts passendes steht, den Korpus-Metadatenblock.

```
<textSigle>WPD/AAA.00001</textSigle>
<t.title assemblage="external">WPD/AAA.00001 Ruru; Jens.Ol;
    Aglarech; u.a.: A, In: Wikipedia 2005</t.title>
<h.title type="main">A</h.title>
<h.author>Ruru; Jens.Ol; Aglarech; u.a.</h.author>
<pubDate type="year">2005</pubDate>
<pubPlace>URL:http://de.wikipedia.org</pubPlace>
<text>
<body>
<div n="0" complete="y" type="Enzyklopädie-Artikel">
<div n="1" complete="y" type="section">
<p><gap desc="Abbildung" reason="omitted"/>
```

```
<s><hi rend="bo">A</hi> bzw. <hi rend="bo">a</hi> ist der erste
    Buchstabe des lateinischen Alphabets und ein Vokal.</s>
<s>Der Buchstabe A hat in deutschen Texten eine
    durchschnittliche Häufigkeit von 6,51 %.</s>
<s>Er ist damit der sechsthäufigste Buchstabe in deutschen
    Texten.</s></p>
```

**Listing 1: Auszug aus dem Wikipedia-Korpus im IDS-XCES-Format**

Nicht alle diese Metadatentypen sollen unmittelbar für die Ad-hoc-Recherche in *KoGra-DB* genutzt werden. Unsere Abfrageoberfläche (siehe hierzu Kapitel 2.6) bietet explizit lediglich die Einbeziehung von Staat (d) und Datum (h) zur Eingrenzung der Suchergebnisse an. Der Texttitel wird im Online-System nicht für die eigentliche Recherche, sondern gemeinsam mit weiteren datenbankintern berechneten Werten wie z.B. der Tokenanzahl als Informationsmerkmal bei der Auflistung der Korpusinhalte verwendet. Als zusätzliche Online-Suchkriterien dienen dagegen folgende von uns teilautomatisch generierte Metadaten:

j) Medium als manuell überprüfte Ableitung aus Texttyp, Genre und Ressort
Ausprägungen: Publikumspresse, Bücher, Gesprochenes, sonstige Printmedien, Internet/Wikipedia

k) Register als manuell überprüfte Ableitung aus Texttyp und Ressort
Ausprägungen: Pressetexte, Gebrauchstexte, Literarische Texte

l) Domäne als thematische Topic-Zusammenfassung
Ausprägungen: Fiktion, Kultur/Unterhaltung, Mensch/Natur, Politik/Wirtschaft/Gesellschaft, Technik/Wissenschaft, unklassifizierbar

m) Geburts- sowie gegebenenfalls Heimatort auf Basis des Publikationsorts und eigener Recherchen

n) Region als Zuweisung aller unterschiedlichen Orte zu einer Großregion
Ausprägungen: Nordwest, Nordost, Mittelwest, Mittelost, Mittelsüd, Südwest (einschließlich Schweiz), Südost (einschließlich Österreich)

Die datenbankinterne Verwaltung der textspezifischen Metadaten dokumentiert Kapitel 2.4.

### 2.2.2 Morpho-syntaktische Annotationen

Bezogen sich die oben aufgelisteten Importdaten ausschließlich auf die Textebene, so beschreiben die morpho-syntaktischen Annotationen nunmehr gleichermaßen Satz-, Phrasen- und Tokenebene. Kombinierte Abfragewünsche bezüglich dieser sowie gegebenenfalls später noch hinzuzufügenden Ebenen unterliegen natur- und erfahrungsgemäß einer kontinuierlichen Entwicklung. Dies impliziert, dass der initiale Import der XML-Inhalte in relationale Strukturen („shredding") keinerlei Informationen vernichten darf, die in einem künftigen Projektabschnitt noch einmal wertvoll sein könnten. Gleichzeitig müssen die Importstrukturen derart angelegt sein, dass sie vielgestaltige Nachführungen und Ergänzungen ermöglichen. Und nicht zuletzt nimmt der Umfang der zu berücksichtigenden Daten massiv zu: Zum Importzeitpunkt umfasste das Referenzkorpus DeReKo ca. 16 Millionen Texte – das bedeutet dieselbe Anzahl von Text-Datensätzen in *KoGra-DB* –, aber bereits ca. 250 Millionen Sätze gemäß Connexor-Analyse und ca. 4,5 Milliarden Connexor-Token. Durch die zusätzliche parallele Annotation mit TreeTagger und dem Xerox-Parser verdreifachen sich die beiden letztgenannten Datenmengen in der *KoGra*-Datenbank. Geringfügige Abweichungen bei der produktspezifischen Bestimmung von Satz- und Tokengrenzen fallen hier kaum ins Gewicht.

All diese Zahlen variieren im Übrigen geringfügig gegenüber den in anderen Kapiteln genannten Quantitäten. Dies liegt in erster Linie daran, dass nicht sämtliche Inhalte des Gesamtbestands in das Untersuchungskorpus aufgenommen wurden. Beispielsweise wurden Dubletten entfernt oder Texte aufgrund unpassender Metadaten ausgeklammert. Außerdem muss beachtet werden, dass die Anzahl der DeReKo-Textwörter von den Zahlen abweicht, die von den darauf operierenden Taggern als Token erkannt werden.

Listing 2 illustriert exemplarisch den standoff-XML-Output des Connexor Machinese Phrase Tagger für die einleitenden Wörter des ersten Satzes aus 2.1. Satz- und Nominalphrasengrenzen werden durch hierarchisch geschachtelte XML-Elemente (`<sentence>` und `<np>`) kodiert. Darin eingebettet finden sich für jedes Token (`<token>`) Angaben zur Position im Primärtext (`pos`), zur Länge (`len`), zum jeweiligen Textwort (`<text>`) und dessen Lemmatisierung (`<lemma>`) sowie zur Wortklasse (`morpho`) und Syntax (`syntax`).[15]

[15] Eine detaillierte Dokumentation des Connexor-Tagsets findet sich unter http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/ (Stand: August 2013).

```
<sentence>
<np>
<token pos="15607" len="1">
<text>A</text>
<lemma>A</lemma>
<tags syntax="@NH" morpho="N"/>
</token>
</np>
<token pos="15614" len="4">
<text>bzw.</text>
<lemma>bzw.</lemma>
<tags syntax="@CC" morpho="CC"/>
</token>
<np>
<token pos="15633" len="1">
<text>a</text>
<lemma>a</lemma>
<tags syntax="@NH" morpho="N"/>
</token>
</np>
<token pos="15640" len="3">
<text>ist</text>
<lemma>sein</lemma>
<tags syntax="@MAIN" morpho="V" sub1="IND" sub2="PRES"/>
</token>
<token pos="15644" len="3">
<text>der</text>
<lemma>der</lemma>
<tags syntax="@PREMOD" morpho="DET"/>
</token>
<np>
<token pos="15648" len="5">
<text>erste</text>
<lemma>erste</lemma>
<tags syntax="@PREMOD" morpho="NUM" sub1="ORD"/>
</token>
<token pos="15654" len="9">
<text>Buchstabe</text>
<lemma>Buchstabe</lemma>
<tags syntax="@NH" morpho="N"/>
</token>
</np>
```

Listing 2: Morpho-syntaktische Connexor-Annotationen

Der initiale Import in die Korpusdatenbank verläuft denkbar unkompliziert. Sämtliche Connexor-Daten – einzig unter Auslassung der Längenangabe, die für unsere Untersuchungsziele absehbar nicht relevant ist bzw. im Bedarfsfall einfach nachzuberechnen wäre – landen zunächst in einer nicht weiter relationierten Basistabelle (TB_CONNEXOR) mit den Spalten CO_POSITION (numerische Position im Primärtext), CO_TOKEN (Textwort), CO_LEMMA (Grundform), CO_SYNTAX (syntaktische Spezifikation), CO_NP (boolscher Indikator der Zugehörigkeit zu einer Nominalphrase), CO_MORPHO (Wortklasse), CO_SUB (zusammengeführte optionale Angaben wie beispielsweise Modus, Tempus oder Aspekt bei Verben). Automatisch hinzugefügt werden fortlaufende eindeutige Nummerierungen der Token (CO_ID) sowie Satznummern (CO_SENTENCEID).

| Wortklasse | Anzahl |
|---|---|
| A | 297 632 979 |
| ADV | 315 742 709 |
| CC | 111 860 775 |
| CS | 48 610 864 |
| DET | 485 748 867 |
| INTERJ | 414 605 |
| N | 1 212 426 943 |
| NUM | 143 721 526 |
| PREP | 412 565 058 |
| PRON | 308 723 388 |
| V | 521 221 803 |
| Satzzeichen | 642 064 741 |
| **TOTAL** | **4 500 734 258** |

Tab. 1: Verteilung der Wortklassen (Connexor) in *KoGra-DB*

Die Basistabelle kann bereits für die Abfrage und Berechnung textübergreifender Kennzahlen genutzt werden; Tabelle 1 stellt beispielsweise die Verteilung der Wortklassen über das Gesamtkorpus dar.

Eine erste Einschätzung der Plausibilität der automatisch generierten Annotationen (siehe hierzu auch Kapitel 4) erlaubt Tabelle 2: Hier wurden sämtliche morphologischen Zusatzangaben aus CO_SUB aufgelistet, gezählt und auf Konformität mit der Connexor-Dokumentation überprüft. Undokumentierte – und unplausible – Annotationen wie die Kombination „Verb im Infinitiv und

| Wortklasse | Zusatz | Anzahl | dokumentiert |
|---|---|---|---|
| A | CMP | 10232971 | j |
| A | CMP PL | 163 | n |
| A | CMP PRES | 827 | n |
| A | IND PL | 29 | n |
| A | IND PRES | 36538 | n |
| A | INF | 102812 | n |
| A | INF PL | 55 | n |
| A | INF PRES | 65209 | n |
| A | PL | 10444 | n |
| A | Prop | 20253 | n |
| A | Prop PL | 395 | n |
| A | SUB PRES | 28663 | n |
| A | SUP | 8430884 | j |
| A | SUP PAST | 69 | n |
| A | SUP PL | 78 | n |
| ADV | PL | 4793 | n |
| N | Abr | 7578367 | j |
| N | Abbr PL | 71300 | j |
| N | CMP PL | 38 | n |
| N | IND PAST | 1 | n |
| N | IND PL | 7858 | n |
| N | IND PRES | 2482 | n |
| N | INF | 1644 | n |
| N | INF PL | 20049 | n |
| N | INF PRES | 1224 | n |
| N | PCP PERF | 195 | n |
| N | PCP PL | 875 | n |
| N | PL | 213627837 | j |
| N | Prop | 293946811 | j |
| N | Prop PL | 1463347 | j |
| N | SUB PL | 495 | n |
| N | SUB PRES | 2009 | n |
| N | SUP PL | 1 | n |
| NUM | ORD | 20805218 | j |
| NUM | Prop | 23 | n |
| PREP | IND PRES | 25 | n |
| PREP | PL | 229 | n |
| PREP | SUB PRES | 11272 | n |
| PRON | INF | 79 | n |
| V | IMP | 2414990 | j |
| V | IND PAST | 110826576 | j |
| V | IND PRES | 210282714 | j |
| V | INF | 68460010 | j |
| V | INF PAST | 683 | j |
| V | INF PERF | 71 | j |
| V | INF PL | 4329 | n |
| V | INF PRES | 6325252 | j |
| V | PCP PERF | 85252139 | j |
| V | PCP PROG | 3023125 | j |
| V | SUB PAST | 12429244 | j |
| V | SUB PRES | 22202670 | j |

Tab. 2: Verteilung der Zusatzangaben (Connexor) in *KoGra-DB*

Plural" (V INF PL) lassen sich auf diese Weise gezielt aufspüren. Gleiches gilt für prinzipiell mögliche, aber aufgrund der (hier: niedrigen) Frequenz auffällige Kombinationen wie „Infinitiv Perfekt".

Ein wenig aufwändiger als das initiale Einlesen der Annotationsdaten gestaltet sich die datenbankinterne Zuordnung der geparsten Sätze zu einem Text. Da sich in den Connexor-Annotationen keine entsprechende Referenz findet, muss für jeden Satz – unter Zuhilfenahme der Position des jeweils ersten Textworts – in der XCES-Datei die zugehörige Textsigle ermittelt werden. Erst mit dieser Information ist später – unabhängig vom konkreten Recherchemodell – eine Eingrenzung der Suchbabfragen auf bestimmte Texte möglich.

## 2.3 Existierende Konzepte für die datenbankbasierte Korpusverwaltung

Die finale Modellierung der ebenenübergreifenden Informationsstrukturen in *KoGra-DB* sowie deren Nutzung für die Recherche werden in den Kapiteln 2.4 bzw. 2.5 dargelegt. Vorab evaluieren wir existierende Konzepte für die datenbankbasierte Korpusverwaltung hinsichtlich ihrer Eignung im gegebenen Forschungskontext. Als Referenzsystem dient ein auf einem typischen Midrange-Server (4-Core-Prozessor mit 2,67 GHz Taktung und 16 GB RAM) implementiertes objekt-relationales Datenbankmanagementsystem (Oracle DBMS); die folgende Evaluation basiert auf vier unterschiedlich umfangreichen Korpora (10 Millionen, 100 Millionen, 1 Milliarde und 4 Milliarden Textwörter). Die Hardwareausstattung bewegt sich bewusst auf einem mittleren Level, da in der Evaluationsphase nur ein sehr eingeschränkter Nutzerkreis zeitgleich recherchieren wird. Unsere Zielsetzung ist die Konzeption eines für den Einbenutzerbetrieb ausreichenden Retrievalsystems, das später unter Produktionsbedingungen auf einem leistungsfähigeren Server mit zusätzlichen Prozessoren und mehr Hauptspeicher optimal skaliert.

### 2.3.1 N-Gram-Tabellen

Davies (2005) stellt am Beispiel des „Corpus del Español"[16] einen relationalen Ansatz für die datenbankbasierte Verwaltung eines linguistischen Korpus mit ca. 100 Millionen Textwörtern sowie deren Lemmata und Wortklassenbestimmungen vor. Der Grundgedanke besteht im Aufbrechen der linearen Korpusstruktur durch die Segmentierung in n-Tupel (geordnete Paare, Tripel,

[16] Das Korpus ist unter http://www.corpusdelespanol.org online abfragbar.

Quadrupel etc.), die anschließend in sogenannten n-Gram-Tabellen abgelegt werden. Die spätere SQL-Recherche erfolgt ausschließlich auf diesen Tabellen bzw. auf den hierauf angelegten Indizes und vermeidet dadurch eine Traversierung kompletter Textbäume.

Abbildung 2 illustriert die entsprechende Tabellenarchitektur für Suchabfragen über einen maximalen Abstand von zwei Textwörtern: Eine 3-Gram-Tabelle speichert jeweils drei aufeinander folgende Textwörter sowie Lemma- und Wortklassenbestimmungen – also drei Attribute pro Textwort – in einer gemeinsamen Tabellenzeile, die durch eine eindeutige ID ansprechbar ist. Insgesamt verteilen sich die Daten auf 10 indizierte Spalten,[17] eine analoge 4-Gram-Tabelle umfasst demnach 13, eine 5-Gram-Tabelle 16 Spalten usw. Jedes tokenspezifische Zusatzattribut erhöht die Spaltenzahl nochmals gemäß der Formel:

3-Gram

| ID | Wort 1 | Lemma 1 | Wortklasse 1 | Wort 2 | Lemma 2 | Wortklasse 2 | Wort 3 | Lemma 3 | Wortklasse 3 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | A | A | N | bzw. | bzw. | CC | a | a | N |
| 2 | bzw. | bzw. | CC | a | a | N | ist | sein | V |
| 3 | a | a | N | ist | sein | V | der | der | DET |
| 4 | ist | sein | V | der | der | DET | erste | erste | NUM |
| 5 | der | der | DET | erste | erste | NUM | Buchstabe | Buchstabe | N |

**Abb. 2: 3-Gram-Tabelle**

Der Vorteil dieser Vorgehensweise liegt darin, dass darauf verzichtet wird, Daten aus unterschiedlichen Tabellen oder Zeilen während der Abfrage zu verknüpfen. Verknüpfungen dieser Art werden in der Fachlieratur als „SQL-Joins" bezeichnet. Ihre Ausführung ist rechenintensiv und dauert folglich zumeist vergleichsweise lange. Um beispielsweise eine Liste sämtlicher durch den Konnektor *bzw.* unmittelbar verknüpften Nomina zu erstellen, wäre folgende SQL-Abfrage zielführend:

```
select Wort1, Wort3 from <3-Gram> where Wortklasse1='N' and
    Wortklasse2='CC' and Lemma2='bzw.' and Wortklasse3='N';
```

**Listing 3: Abfrage der 3-Gram-Tabelle**

Der Praxistest auf unserem Referenzsystem ergab erwartungsgemäß durchgehend kurze Abfragezeiten von unter einer Sekunde (10 Millionen-Korpus) bis maximal 10 Sekunden (4-Milliarden-Korpus) mit signifikanten, aber letztlich tolerierbaren Abweichungen je nachdem, ob mit Einzelspalten- oder Mehr-

[17] Davies ergänzt diese Architektur noch um separate Spalten mit Frequenzangaben, was für unsere Anfragen jedoch nicht erforderlich ist und keinerlei Auswirkungen auf die Abfragezeiten hat.

spaltenindizes (also mit separaten Indizes für einzelne Abfrageattribute oder kombinierten Indizes für mehrere Abfrageattribute) gearbeitet wurde.

Kritisch anzumerken bleibt die dem Modell inhärente Redundanz der physischen Datenhaltung: Durch die Mehrfachspeicherung aller erdenklichen Abfrageattribute erhöht sich bereits bei 3-Gram-Tabellen ohne Indizes die vorzuhaltende Datenmenge um den Faktor 3, das Hinzufügen von Indizes erfordert darüber hinaus beträchtlichen Speicherplatz.[18] Sollen weiterhin auch größere Wortabstände recherchierbar gemacht werden, erfordert das Konzept sehr rasch 10- oder 20-Gram-Tabellen mit entsprechenden Volumenanforderungen und kann letztlich doch nicht verhindern, dass Suchmuster mit noch größeren internen Abständen – beispielsweise für satzübergreifende grammatische Phänomene – nicht abgedeckt werden.

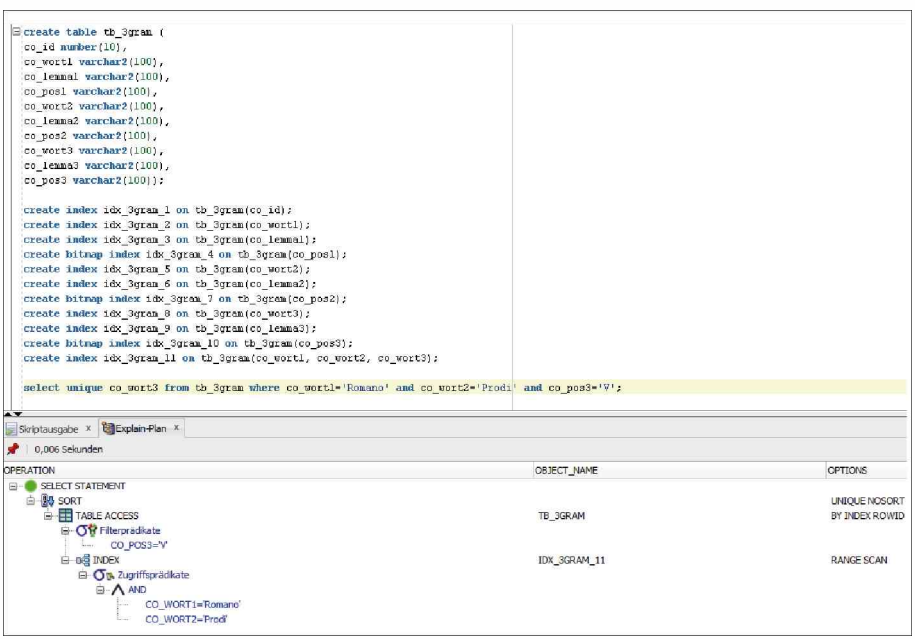



**Abb. 3: Ausführungsplan einer SQL-Abfrage der 3-Gram-Tabelle**

[18] Angesichts der Tatsache, dass Abfragen in der Regel nicht auf den eigentlichen Tabellendaten, sondern auf Indizes durchgeführt werden, wäre in Abweichung von Davies' Konzept der Einsatz sogenannter „index-organized tables" konsequent. Diese speichern Datenzeilen nicht unsortiert, sondern direkt als Baumindex und benötigen folglich keine zusätzlichen Indizes. Eine weitere Alternative wären funktionale Indizes auf 1-Gram-Tabellen, welche die n-Tupel nicht in Spalten, sondern im Index ablegen. Das Grundproblem der redundanten Speicherung bleibt allerdings in beiden Fällen bestehen.

Abgesehen von Volumenaspekten, die angesichts heutzutage moderater Anschaffungskosten für Festplattenspeicher zumindest aus finanzieller Sicht nicht zwangsläufig zum Ausschluss dieses Ansatzes führen müssen, geht der Anstieg der Spaltenzahl in vielen Fällen leider mit einer Verlängerung der Abfragezeiten einher. Die logische Begründung hierfür erschließt sich leicht, wenn man den Ausführungsplan einer Abfrage mit mehreren nicht-hochfrequenten Suchausdrücken – also beispielsweise nicht-„top-freq"-Token oder -Lemmata[19] – betrachtet: Abbildung 3 demonstriert, dass der kostenbasierte Optimierer („Cost-Based-Optimizer" bzw. „CBO") der Datenbank in diesen Fällen bei Vorhandensein von sowohl Einzel- und Mehrspaltenindizes die zusammengesetzte Variante (hier: Index idx_3gram_11) bevorzugt.[20] Deren Struktur jedoch expandiert analog zur Erhöhung der Spaltenzahl. Die Selektivität eines einzelnen Spaltenwerts nimmt dadurch ab und führt in der Folge tendenziell zu höheren Suchkosten.

Mindestens ebenso schwer wiegt der Umstand, dass eine SQL-Abfrage von n-Gram-Tabellen überhaupt nur dann performant, d.h. unter Ausnutzung von Indizes, durchführbar ist, wenn keine NOT-Muster (also beispielsweise „Finde alle Nomina mit Wortabstand 2, die *nicht* durch den Konnektor *bzw.* verknüpft sind") zum Einsatz kommen; in diesen Fällen wäre ein teurer (= langwieriger) „full table scan" notwendig, d.h. die sequentielle Abarbeitung aller Tabellenzeilen. Auf unserem Referenzsystem benötigt eine solche Abfrage bereits ca. 10 Minuten für das 4-Milliarden-Korpus. Gleiches gilt im Übrigen für dynamische Recherchen mit regulären Ausdrücken: Auch hier bieten traditionelle Baum- und Bitmap-Indizes keine Hilfestellung, weil das zur Abfragezeit konkret eingegebene Suchmuster angesichts der Vielzahl potenziell möglicher Varianten vorab gar nicht indizierbar ist.[21]

### 2.3.2 Relationierung und Abfrage mit SQL-Joins

Eine alternative Strategie basiert auf der vollständigen Relationierung aller Primär- und Annotationsdaten sowie auf der Übersetzung von Abfragen in geschachtelte SQL-Joins, d.h. in durch logische Operatoren wie „AND" verbundene Unterabfragen. Chiarcos et al. (2008) verfolgen diesen Ansatz für die

[19] Siehe hierzu die Unterteilung der Textwörter in Frequenzklassen im folgenden Anschnitt.

[20] Für Spalten mit relativer hoher Kardinalität wie z.B. Wortklassenangaben bieten sich dagegen sogenannte „Bitmap-Indizes" auf Einzelspalten an.

[21] Mittelfristig könnten hier neuere Ansätze für das Indizieren regulärer Ausdrücke Abhilfe schaffen; vgl. z.B. Majumder et al. (2008). Unseres Wissens existieren allerdings noch keine produktiven Implementierungen für professionelle Datenbanksysteme.

linguistische Datenbank ANNIS; Bird et al. (2005) messen die Abfragezeiten eines entsprechenden Datenbank-Frameworks für mehrere – allerdings vergleichsweise kleine – Testkorpora. Dabei wird ein Annotationsbaum zur Ladezeit geparst und jeder Knoten mit Angaben zu den jeweiligen Vater- und Geschwisterknoten in einer separaten Tabellenzeile abgelegt.

Abbildung 4 veranschaulicht diese Vorgehensweise für unsere bereits in Abbildung 2 verwendeten Beispieldaten. Da jede Angabe nur einmal gespeichert wird, erhöht sich das Datenvolumen nicht unnötig. Jede Tabellenspalte erhält einen passenden B-Baum- bzw. Bitmap-Index.[22] Durch die Verwendung fortlaufender Satz- und Token-IDs entfällt die Limitierung der Wortabstände in Suchabfragen. Listing 4 illustriert eine geschachtelte SQL-Query, die analog zu Listing 3 Wortklassen- und Lemmaattribute für das Auffinden sämtlicher durch den Konnektor *bzw.* unmittelbar verknüpfter Nomina heranzieht. Jedes tokenspezifische Suchattribut wird dabei auf einen SQL-Join abgebildet.

RelTab

| ID | Satz-ID | Wort | Lemma | Wortklasse |
|---|---|---|---|---|
| 2 | 1 | A | A | N |
| 3 | 1 | bzw. | bzw. | CC |
| 4 | 1 | a | a | N |
| 5 | 1 | ist | sein | V |
| 6 | 1 | der | der | DET |
| 7 | 1 | erste | erste | NUM |
| 8 | 1 | Buchstabe | Buchstabe | N |

**Abb. 4: Relationale Speicherung der Primär- und Annotationsdaten**

```
select t1.Wort, t3.Wort from <RelTab> t1, (select * from <RelTab>
where Wortklasse='CC' and Lemma='bzw.') t2, (select * from
<RelTab> where Wortklasse='N') t3 where t1.Wortklasse='N' and
t1.Satz-ID = t2.Satz-ID and t1.Satz-ID = t3.Satz-ID and
t1.ID+1>t2.ID and t2.ID+1>t3.ID;
```

**Listing 4: Abfrage der relationierten Annotationen mit SQL-Joins**

Die gleichermaßen übersichtliche wie elegante Relationierung besticht durch einen vergleichsweise geringen Einlese- und Indizierungsaufwand sowie die Nutzung etablierter Indexstrukturen für Abfragen mit beliebigen Wortabständen. Unsere in Kapitel 2.2 eingeführte Basistabelle lässt sich beispielsweise

[22] Zur Verwendung dieser Indexvarianten siehe z.B. die Testreihe von Vivek Sharma unter http://www.oracle.com/technetwork/articles/sharma-indexes-093638.html (Stand: 2.8.2013).

ohne weitere Änderungen für die Recherche nutzen. Dessen ungeachtet gilt es zu prüfen, welche Auswirkungen der Einsatz geschachtelter SQL-Joins für sehr große Datenmengen und unterschiedlich frequente Suchattribute auf die absoluten Retrievalzeiten hat. Um einen empirischen Nachweis zu führen, verwenden wir wieder unsere indizierten Testkorpora mit 10 Millionen, 100 Millionen, 1 Milliarde und 4 Milliarden Textwörtern. Weiterhin teilen wir die enthaltenen Textwörter in jeweils eine von fünf Frequenzklassen ein:

- „rare": Frequenz < 25 000 im 4-Milliarden-Korpus; Beispiele: *traurige* (18 538), *isoliert* (18 401), *trauen* (23 884)
- „low-freq": Frequenz < 150 000 im 4-Milliarden-Korpus; Beispiele: *langsam* (140 303), *lesen* (139 896), *verfügt* (107 985)
- „mid-freq": Frequenz < 2 000 000 im 4-Milliarden-Korpus; Beispiele: *ohne* (1 686 945), *nun* (1 996 277), *uns* (1 928 430)
- „high-freq": Frequenz < 25 000 000 im 4-Milliarden-Korpus; Beispiele: *nicht* (17 750 063), *ist* (20 013 587), *dem* (18 509 294)
- „top-freq": Höchste Frequenz im 4-Milliarden-Korpus; Beispiele: *der* (90 822 004), *die* (82 726 036), *und* (63 804 894)

Die aufgeführten Beispiel-Textwörter können nun zur Formulierung frequenzklassenspezifischer SQL-Abfragen eingesetzt werden. Dabei beschränken wir uns auf das reine Zählen der Vorkommen vermittels des `count`-Operators und legen fest, dass die Textwörter einer Klasse in einem gemeinsamen Satz nacheinander vorkommen sollen:

1) Recherche mit einem Suchattribut und ohne SQL-Join: `select count (t1. sentenceid) from <tokenTab> where token =` umfangreichen `<token>`; als `<tokenTab>` setzen wir nacheinander die unterschiedlich Korpustabellen ein, `<token>` steht für das Beispiel-Textwort einer Frequenzklasse.
2) Recherche mit zwei Suchattributen und einem SQL-Join: `select count(t1.sentenceid) from <tokenTab> t1, (select id, sentenceid from <tokenTab> where token = <token2>) t2 where token=<token1> and t1.sentenceid = t2.sentenceid and t2.id>t1.id;` `<token1>` und `<token2>` stehen für zwei verschiedene Textwörter einer Frequenzklasse.

3) Recherche mit drei Suchattributen und zwei SQL-Joins: `select count(t1.sentence_id) from <tokenTab> t1, (select id, sentence_id from <tokenTab> where token = <token2>) t2, (select id, sentence_id from <tokenTab> where token = <token3>) t3 where token = <token1> and t1.sentence_id = t2.sentence_id and t1.sentence_id = t3.sentence_id and t3.id>t2.id and t2.id>t1.id;` *<token1>*, *<token2>* und *<token3>* stehen für drei verschiedene Textwörter einer Frequenzklasse.

Die Abbildungen 5 bis 7 zeigen unsere ermittelten Suchzeiten in Sekunden für die wechselnden Frequenzklassen und Korpusgrößen. Jede Abfrage wurde zur Bestimmung eines aussagekräftigen Durchschnittswerts dreimal initiiert. Um Caching-Effekte – d.h. die Nutzung von im (schnellen) Hauptspeicher des Abfrageservers noch von einer vorherigen Abfrage vorhandenen Ergebnissen – auszuschließen, wurden die entsprechenden DBMS-internen Speicherstrukturen vor jeder Abfrage geleert.

| | 1 Mio | 10 Mio | 100 Mio | 1 000 Mio | 4 000 Mio |
|---|---|---|---|---|---|
| **rare** (*traurige*) | 0,04 | 0,05 | 0,1 | 0,2 | 0,2 |
| **low** (*langsam*) | 0,03 | 0,05 | 0,2 | 0,3 | 0,3 |
| **mid** (*ohne*) | 0,1 | 0,1 | 0,2 | 0,3 | 0,3 |
| **high** (*nicht*) | 0,21 | 0,28 | 0,45 | 1 | 1 |
| **top** (*der*) | 0,29 | 0,49 | 0,92 | 5 | 6 |

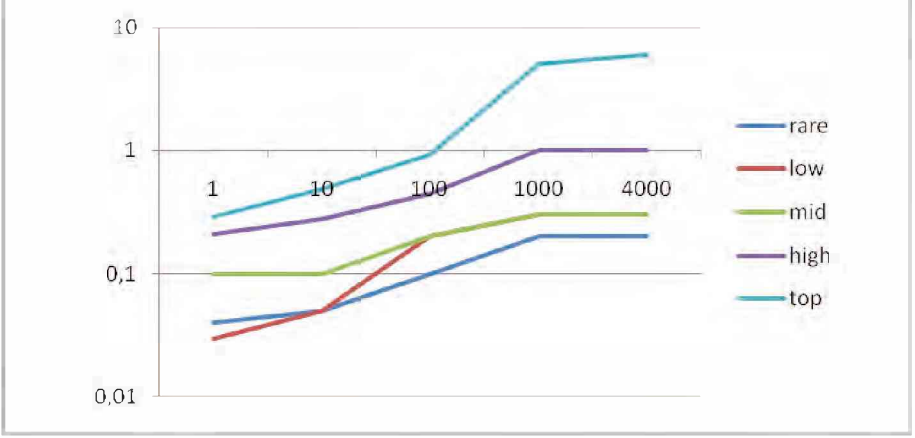



Abb. 5: Abfragezeiten (in s) für ein Suchattribut (logarithmische Darstellung)

| | 1 Mio | 10 Mio | 100 Mio | 1 000 Mio | 4 000 Mio |
|---|---|---|---|---|---|
| **rare** (*traurige/isoliert*) | 0,16 | 0,19 | 0,29 | 0,3 | 0,5 |
| **low** (*langsam/lesen*) | 0,09 | 0,2 | 0,3 | 0,4 | 1,3 |
| **mid** (*ohne/nun*) | 0,16 | 0,23 | 0,31 | 0,9 | 2,3 |
| **high** (*nicht/ist*) | 0,21 | 0,32 | 0,89 | 6,8 | 12,7 |
| **top** (*der/die*) | 0,32 | 0,65 | 3,12 | 56 | 63,3 |

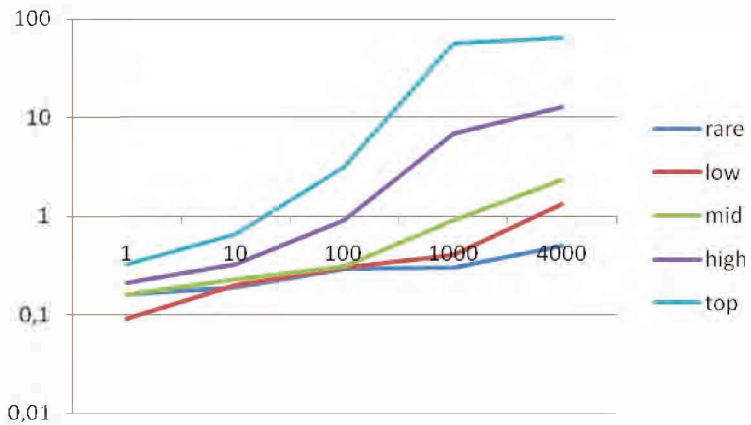



Abb. 6: Abfragezeiten (in s) für zwei Suchattribute (logarithmische Darstellung)

| | 1 Mio | 10 Mio | 100 Mio | 1 000 Mio | 4 000 Mio |
|---|---|---|---|---|---|
| **rare** (*traurige/isoliert/trauen*) | 0,03 | 0,16 | 0,3 | 0,6 | 0,9 |
| **low** (*langsam/lesen/verfügt*) | 0,29 | 0,35 | 0,37 | 0,65 | 1 |
| **mid** (*ohne/nun/uns*) | 0,3 | 0,4 | 0,5 | 1,2 | 3,8 |
| **high** (*nicht/ist/dem*) | 0,31 | 0,47 | 1,49 | 10,47 | 38,7 |
| **top** (*der/die/und*) | 0,4 | 0,87 | 4,67 | 47,06 | 301 |

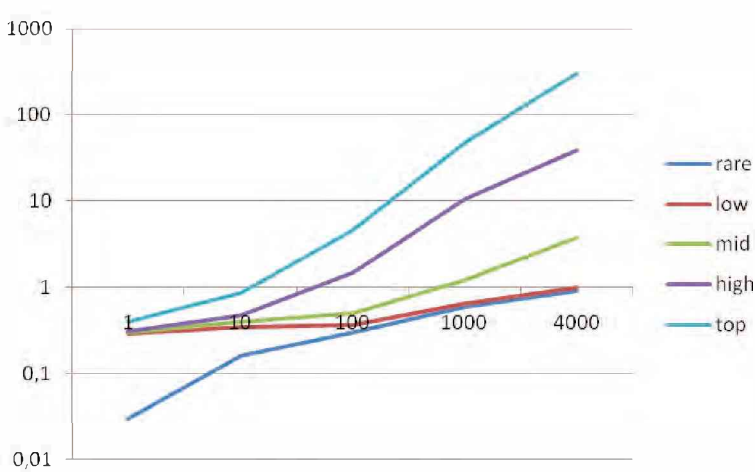



Abb. 7: Abfragezeiten (in s) für drei Suchattribute (logarithmische Darstellung)

Es wird deutlich, dass die Abfragezeiten für steigende Korpusgrößen und die Anzahl der Suchattribute rasch den Rahmen übersteigen, der für ein Ad-hoc-Retrievalsystem annehmbar ist. Die SQL-Joins generieren, trotz optimaler Nutzung der Indizes, eine derart hohe Arbeitslast, dass bereits Testläufe mit drei Suchwerten auf unserem Referensystem teilweise mehrere Minuten benötigen. Insbesondere skalieren die Zeiten nicht linear, wie folgendes Beispiel illustriert: Während die Suche nach einem einzelnen „top-freq"-Token im 4-Milliarden-Korpus noch 6 Sekunden und die Suche nach zwei aufeinander folgenden Textwörtern 63 Sekunden dauert, müssen wir auf das Ergebnis der Recherche nach drei aufeinander folgenden „top-freq"-Textwörtern bereits über 5 Minuten warten.

In Anbetracht der Tatsache, dass reale Korpusanfragen in den allermeisten Fällen zusätzliche, satzübergreifende Suchkriterien beinhalten – und zwar gleichermaßen Kriterien für benachbarte Sätze (beispielsweise zur Recherche nach anaphorischen Phänomenen) wie auch „übergeordnete" textspezifische Metadaten; vgl. hierzu Kapitel 2.2.1 – und dadurch zusätzliche SQL-Joins notwendig machen, erscheint diese Suchstrategie für unsere Zwecke als unzweckmäßig; kombinierte Testläufe mit drei und mehr SQL-Joins auf den Text- und Token-Tabellen[23] mussten wir teilweise nach mehreren Stunden abbrechen. Als Fazit bleibt daher die Erkenntnis, dass die Relationierung zwar zu einer übersichtlichen und handhabbaren Datenhaltung führt, komplexe Mehrebenen-Recherchen jedoch auf ein besser skalierbares Retrievalkonzept angewiesen sind.

## 2.4 Das *KoGra-DB*-Datenmodell

Ausgehend von der oben begründeten Entscheidung, die Korpusdaten weitestgehend relationiert – und nicht z.B. als n-Gramme – zu speichern, erfolgt nun die semantische Datenmodellierung. Dabei geht es darum, den zu erfassenden Weltausschnitt, d.h. die in Kapitel 2.2 aufgeführten Korpusinhalte, mit Hilfe eines theoretischen Instrumentariums formal so abzubilden, dass sich daraus anschließend die benötigten physikalischen Tabellenstrukturen ableiten lassen. Wir folgen dabei der Notation von Peter Chen, dem Begründer der Entity-Relationship-Diagramme (ER-Diagramme).[24] Abbildung 8 illustriert

[23] Wobei die Tokentabellen wie oben beschrieben für die Speicherung von tokenspezifischen Attributen und die Texttabellen für die Speicherung von textspezifischen Metadaten verwendet wurden.

[24] Vgl. Chen (1976).

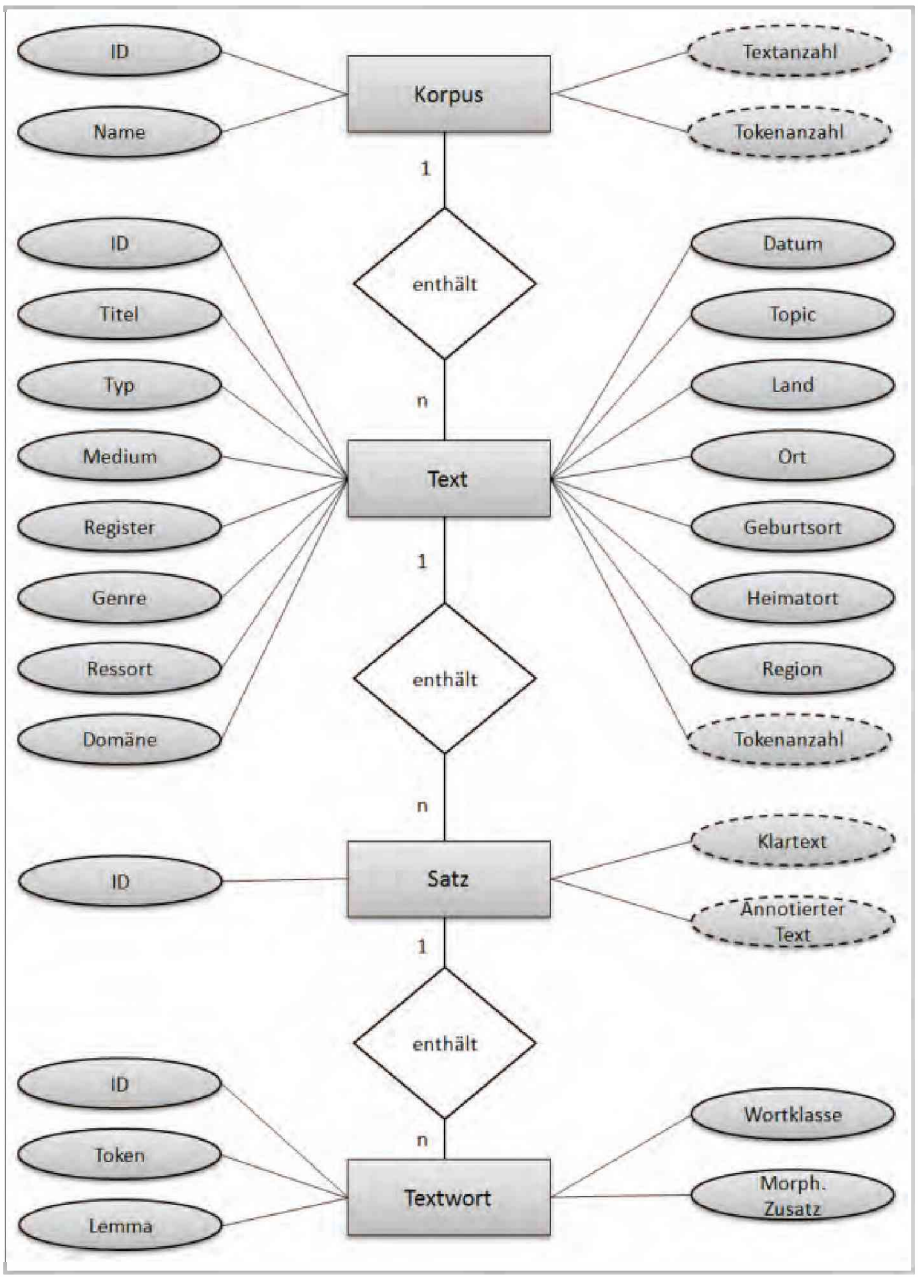



Abb. 8: Semantisches Datenmodell *KoGra-DB*

unsere Modellierung: Die erfassten Entitätstypen (Korpus, Text, Satz, Textwort) werden durch Rechtecke, deren Attribute in Ellipsenform dargestellt; abgeleitete bzw. berechnete Attribute sind durch punktierte Ränder gekennzeichnet. Beziehungstypen, die Zusammenhänge zwischen den Entitätstypen ausdrücken, werden als Rauten abgebildet. Die an den Verbindungslinien zwischen den Entitätstypen ergänzten Zahlen zeigen an, wie oft eine Entität an einer Beziehung teilnehmen kann, wobei N für „unbegrenzt" steht. Ein Korpus besteht also beispielsweise aus grundsätzlich beliebig vielen Texten, ein Text aus beliebig vielen Sätzen usw.

Erklärungsbedürftig sind die abgeleiteten Attribute: „Textanzahl" (für Korpus) und „Tokenanzahl" (jeweils für Korpus und Text) bilden keine in den Quelldaten vorhandenen Objekte ab, sondern sollen nach erfolgtem Datenimport einmal gezählt und anschließend dauerhaft vorgehalten werden. Diese Konzeption trägt dem Umstand Rechnung, dass solche Angaben erfahrungsgemäß eine häufig genutzte Grundlage für spätere statistische Auswertungen sind. Ähnlich verhält es sich mit den Satz-Attributen „Klartext" und „Annotierter Text". Ersteres bildet eine Konkatenation sämtlicher Textwörter eines Satzes, Letzteres dessen mit Hilfe von XML annotierte Mikrostuktur ab. Die Anwendungsszenarien sind vielschichtig: Beide können beispielsweise dazu eingesetzt werden, um dem Benutzer nach erfolgter Recherche – die im Normalfall eine Liste von Satz-IDs liefert – ohne erneuten Zugriff auf die umfangreiche Tokensammlung ein textuelles bzw. annotiertes Ergebnis zu präsentieren. Weiterhin bieten sie eine Grundlage für die Evaluation alternativer Retrievalstrategien: Enthält eine Suchanfrage keine Annotationsinhalte (Lemma, Wortklasse etc.), so bietet sich die Nutzung eines auf dem Klartext basierenden Volltextindex an. Der annotierte Text lässt sich für die Erprobung der in unserem DBMS integrierten XML-Suchfunktionalitäten[25] nutzen.

Um unseren semantischen Entwurf im relationalen DBMS zu implementieren, überführen wir ihn zunächst in ein physikalisches Datenmodell (siehe Abbildung 9). Dieses konkretisiert die notwendigen Tabellenstrukturen sowie die zur Realisierung von Abhängigkeiten erforderlichen Fremdschlüssel („foreign keys", im Diagramm als Verbindungslinien zwischen den Tabellen kodiert). Ein Primärindex (eindeutiges Schlüsselfeld) wird durch eine Raute (#) gekennzeichnet, obligatorisch gefüllte Tabellenspalten durch ein Sternchen (*) und optional leere Spalten durch eine Null (0).

[25] Vgl. Schneider (2009).

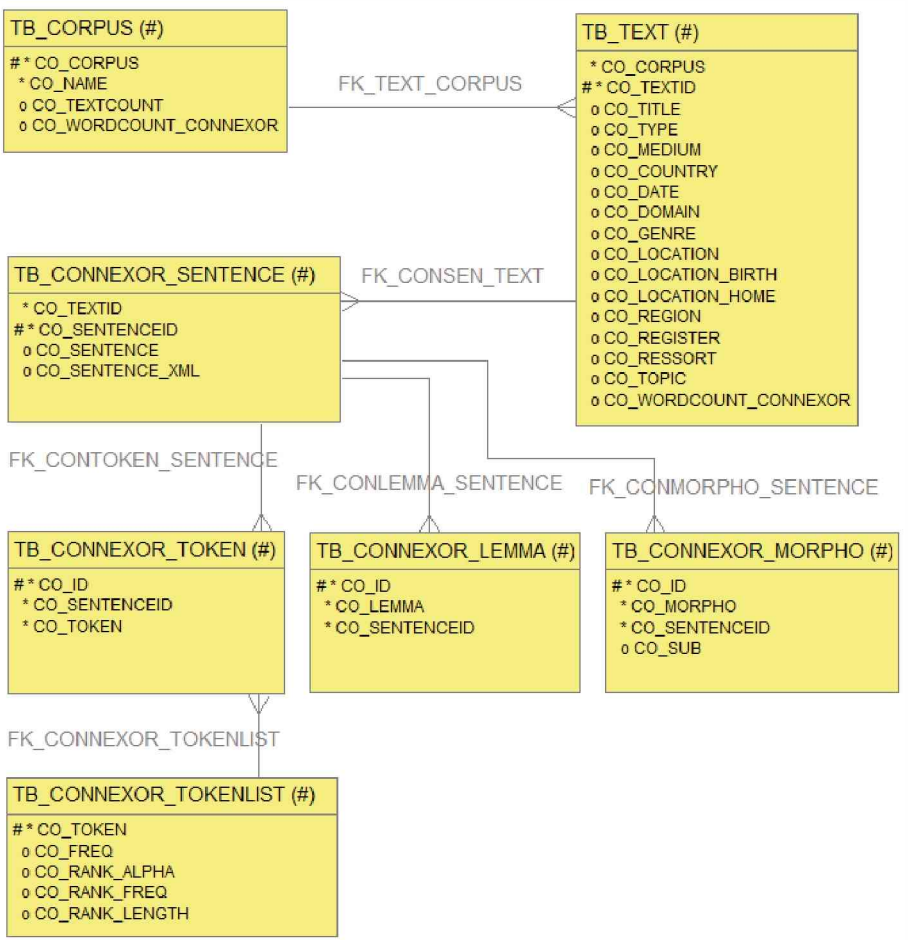



Abb. 9: Physikalisches Datenmodell *KoGra-DB*

Aus Platzgründen gibt unsere Darstellung nur einen repräsentativen Ausschnitt des Gesamtmodells wieder. Ausgeklammert bleiben, wie bereits beim semantischen Entwurf, die Tabellen für die TreeTagger- und Xerox-Annotationen (die analog zur Connexor-Ebene modelliert und implementiert sind). Weiterhin fehlen einige temporäre Tabellen zur Speicherung von Zwischen- und Endergebnissen während der Suche, auf die in Kapitel 2.5 genauer eingegangen wird. Enthalten ist dagegen TB_CONNEXOR_TOKENLIST, eine klassische Lookup-Tabelle zur Speicherung aufwändiger Berechnungen, die sämtliche vom Connexor-Parser ermittelten Token/Types sowie deren Frequenz im Gesamtkorpus und verschiedene Rangpositionen (alphabetisch,

nach Häufigkeit, nach Wortlänge) aufnimmt. Diese Daten sollen in unserer Rechercheumgebung bei Bedarf als geordnete Listen abrufbar sein, siehe hierzu Kapitel 2.6.1.

Erwähnenswert ist weiterhin, dass mehrere textspezifische Metadaten in TB_Text nicht ausgeschrieben, sondern als numerische Werte erscheinen. Die Domänen-Ausprägung 'Kultur/Unterhaltung' in Spalte CO_DOMAIN erhält also beispielsweise den Wert „2", was insgesamt zu einer minimalen – aber dennoch messbaren – Beschleunigung späterer Abfragen führt.

Bevor die im physikalischen Datenmodell dokumentierten Tabellenstrukturen per SQL-DDL eingerichtet und per SQL-DML befüllt werden,[26] stehen noch weitere anwendungsabhängige Design-Entscheidungen an. Diese lassen sich am Beispiel der Morphologie-Tabelle (TB_CONNEXOR_MORPHO) veranschaulichen:

a) Sollen die Daten komprimiert oder unkomprimiert vorgehalten werden? Hier gilt es abzuwägen, ob die durch eine transparente Kompression entstehende zusätzliche Rechenlast durch verbesserte Weiterverarbeitungsoptionen aufgewogen wird. Komprimierte Daten senken im Allgemeinen die Zahl der notwendigen Festplattenzugriffe während der Abfrage, weil damit mehr Datensätze in einen Datenblock passen, und verbessern das Caching-Verhalten, weil auch der Datenbank-Puffer mehr Datensätze aufnehmen kann. Umfangreiche Testreihen auf unserem Referenzsystem ergaben eine geringfügige Beschleunigung typischer Abfragen durch Datenkomprimierung.

b) Soll eine Abfrage-Parallelisierung auf Tabellenebene ermöglicht werden? Da unser Retrievalmodell bei Bedarf das parallele Abarbeiten einzelner Abfragemuster vorsieht, sehen wir auf unserem Quadcore-Server einen maximalen Parallelisierungsgrad von 4 als angemessen an.

c) Sollen Änderungen am Datenbestand (INSERT, UPDATE, DELETE) in Logdateien festgehalten werden, um eine gegebenenfalls notwendiges Wiederherstellen (ROLLBACK) zu unterstützen? Da unsere Korpusinhalte voraussichtlich nur einmal eingelesen und danach nicht mehr verändert werden sollen (die meisten Tabellen werden deshalb sogar mit Schreib-

[26] Die im SQL-Standard enthaltenen Befehle werden üblicherweise in die Kategorien DDL (Data Definition Language) zur Implementierung der Datenbankobjekte und DML (Data Manipulation Language) zur Datenmanipulation unterteilt; daneben existieren noch DCL (Data Control Language)-Befehle für die Rechteverwaltung.

schutz angelegt), entscheiden wir uns für den Verzicht auf Logging-Unterstützung. Dadurch reduziert sich insbesondere bei sehr umfangreichen Tabellen die Ausführungszeit von Einfügungen und Abfragen erheblich.

d) Soll die Tabelle partitioniert werden? Das Konzept der Partitionierung ermöglicht die Aufteilung der Tabelleninhalte unter Verwendung eines Partitionsschlüssels; unser DBMS erlaubt darüber hinaus auch den Einsatz von Unterpartitionen. Für die spätere Abfrage bedeutet dies, dass diejenigen Partitionen ignoriert werden können, in denen der konkrete Suchwert sicher nicht vorkommt. Weiterhin kann die Suche auf verschiedenen Partitionen parallel stattfinden. Die Morphologie-Tabelle wurde deshalb – ebenso wie die Lemma- und Token-Tabelle – anhand der Satznummer (CO_SENTENCEID) in 13 primäre Partitionen aufgeteilt („range partitioning"). Jede Partition nimmt dabei ca. 20 Millionen Tabellenzeilen auf; diesen Wert haben wir durch Versuche mit niedrigeren/höheren Range-Werten als für unser System optimal bestimmt. Beim Anlegen von Unterpartitionen orientieren wir uns an den möglichen Wortklasse-Werten in der Spalte CO_MORPHO („list partitioning"), so dass jede primäre Partition nochmals in 12 Unterpartitionen aufgeteilt wird. Insgesamt enthält die Morphologie-Tabelle folglich 156 Partitionen.

Listing 5 veranschaulicht das Anlegen und Befüllen der Morphologie-Tabelle durch ein „CREATE TABLE AS SELECT (CTAS)" unter Einbeziehung der Basistabelle TB_CONNEXOR. Die Parameter „compress", „parallel 4", „nologging" und „partition by" bzw. „subpartition by" implementieren die oben aufgeführten Design-Entscheidungen.

```
create table tb_connexor_morpho (co_sentenceid, co_morpho,
co_sub, co_id) compress parallel 4 nologging partition
by range(co_sentenceid) subpartition by list (co_morpho)
subpartition template (subpartition A values ('A'),
subpartition V values ('V'), subpartition N values ('N'),
subpartition NUM values ('NUM'), subpartition ADV values
('ADV'), subpartition PRON values ('PRON'), subpartition
DET values ('DET'), subpartition CS values ('CS'),
subpartition CC values ('CC'), subpartition PREP values
('PREP'), subpartition INTERJ values ('INTERJ'),
subpartition P values ('P')) (partition p1 values less
than (20000000), partition p2 values less than (40000000),
partition p3 values less than (60000000), partition p4
values less than (80000000), partition p5 values less than
```

```
(100000000), partition p6 values less than (120000000),
partition p7 values less than (140000000), partition p8
values less than (160000000), partition p9 values less than
(180000000), partition p10 values less than (200000000),
partition p11 values less than (220000000), partition p12
values less than (240000000), partition p13 values less than
(MAXVALUE)) as select co_sentenceid,co_morpho,co_sub,co_id
from tb_connexor;
```

**Listing 5: Anlegen und Füllen der morphologischen Tabelle (Connexor)**

Die Tabellen 3 und 4 dokumentieren den Wertebereich der einzelnen Partitionen; eine Suche nach sämtlichen Korpussätzen mit einer Satznummer bis einschließlich 20 000 000 und mindestens einem enthaltenen Pronomen kann sich dadurch beispielsweise auf die Unterpartition P1_PRON beschränken. Anzumerken bleibt, dass die auf den Tabellen angelegten Indizes identische logische Partitionierungsattribute verwenden („equipartitioned indexes"), so dass nicht nur sequentielle Tabellensuchen („full table scans"), sondern auch Index-basierte Abfragen unsere Aufteilung nutzen können.

| Partitionsname | Hoher Wert CO_SENTENCEID | Unterpartitionen |
|---|---|---|
| P1 | 20 000 000 | 12 |
| P2 | 40 000 000 | 12 |
| P3 | 60 000 000 | 12 |
| P4 | 80 000 000 | 12 |
| P5 | 100 000 000 | 12 |
| P6 | 120 000 000 | 12 |
| P7 | 140 000 000 | 12 |
| P8 | 160 000 000 | 12 |
| P9 | 180 000 000 | 12 |
| P10 | 200 000 000 | 12 |
| P11 | 220 000 000 | 12 |
| P12 | 240 000 000 | 12 |
| P13 | MAXVALUE | 12 |

**Tab. 3: Primäre Partitionierung der morphologischen Tabelle**

| Partitionsname | Unterpartitionsname | Werte CO_MORPHO |
|---|---|---|
| P1 | P1_A | 'A' |
| | P1_V | 'V' |
| | P1_N | 'N' |
| | P1_NUM | 'NUM' |
| | P1_ADV | 'ADV' |
| | P1_PRON | 'PRON' |
| | P1_DET | 'DET' |
| | P1_CS | 'CS' |
| | P1_CC | 'CC' |
| | P1_PREP | 'PREP' |
| | P1_INTERJ | 'INTERJ' |
| | P1_P | 'P' |

**Tab. 4: Subpartitionierung der morphologischen Tabelle (Auszug)**

Einzelne und kombinierte Abfragen nach Wortklassen profitieren enorm von der gewählten Partitionierungsmethode. So benötigt das SQL-Statement `select count(unique co_sentenceid) from tb_connexor_morpho where co_morpho = 'PRON'` (also das Zählen aller Sätze, in denen ein Pronomen auftaucht) auf unserer Morphologie-Tabelle lediglich ca. 7 Sekunden, während die gleiche Abfrage auf einem unpartitionierten Pendant den Nutzer bereits ca. 40 Sekunden warten lässt. Die Join-Abfrage `select count(unique t1.co_sentenceid) from tb_connexor_morpho t1, tb_connexor_morpho t2 where t1.co_sentenceid = t2.co_sentenceid and t1.co_morpho = 'PRON' and t2.co_morpho='DET'` (also das Zählen aller Sätze, in den ein Pronomen sowie ein Artikel vorkommen) terminiert auf der partionierten Tabelle nach ca. 50 Sekunden, auf dem unpartitionierten Gegenstück erst nach ca. drei Minuten. Bei diesen Resultaten sind selbstverständlich weniger die absoluten Werte als vielmehr die dadurch unter gleichen Rahmenbedingungen ermittelten Tendenzen interessant; in wechselnden Einsatzszenarios beeinflussen noch eine Reihe weiterer Datenbankparameter die Abfragezeit.

## 2.5 Abfrage der *KoGra-DB*

Ausgehend von der in Kapitel 2.3 begründeten Annahme, dass trotz der grundsätzlichen Eignung des relationalen Ansatzes für die Pflege sehr umfangreicher annotierter Korpusdaten eine 1:1-Übersetzung komplexer Suchmuster in SQL-Joins nicht ausreichend skaliert, setzen wir als Retrievalmodell ein neuartiges funktionales Verfahren ein. Dabei wird der nicht nur in der Informatik wohlbekannte Grundsatz, dass bei der Zerlegung eines Problems in kleinere Teilprobleme zumeist der Lösungsaufwand sinkt, aufgegriffen und für linguistisch motivierte Korpusdatenbanken empirisch nachgewiesen.

### 2.5.1 Abfragen auf Wortebene

Unser Retrievalmodell orientiert sich am 2004 von Google eingeführten und 2010 als Framework patentierten MapReduce-Ansatz.[27] Dessen erklärte Strategie liegt in der Aufteilung (*mapping*) vielschichtiger Abfragen in performant berechenbare Einzelteile und deren parallelen Abarbeitung auf potenziell ebenfalls verteilten Datenclustern. Auf diese Weise soll der laufzeitverlängernde „Flaschenhals" überwunden werden, der beispielsweise dann entsteht, wenn eine komplexe SQL-Abfrage diverse Tabellen-Joins initiiert. Die verschiedene Suchwerte werden dabei auf sogenannte Map-Prozesse verteilt und deren Ergebnisse in temporären Datenstrukturen gespeichert. Anschließend werden diese Ergebnisse in gegebenenfalls mehreren Reduce-Phasen gefiltert und zusammengeführt. Dieses Vorgehen ist zwar nicht grundsätzlich neu, sondern beruft sich auf die in funktionalen Programmiersprachen wie Lisp bereits etablierten map/reduce-Funktionen, eröffnet uns aber im Kontext des datenbankgestützten Online-Retrievals umfangreicher Mehrebenen-Korpora bislang kaum genutzte Möglichkeiten.

Abbildung 10 illustriert den MapReduce-Datenfluss für eine kombinierte Korpusrecherche mit sieben einzelnen Suchmustern, die sich auf unterschiedliche Annotationsebenen beziehen. Natürlichsprachig formuliert lautet die Beispielabfrage: „Finde alle Sätze, die das Token *mit* enthalten (Token=*mit* im ersten grauen Kasten links oben), das unmittelbar von einem auf *em* endenden (Token=**em*) Adjektiv (POS=A) gefolgt wird, welches seinerseits unmittelbar von einem Adjektiv (POS=A) mit der Endung *en* (Token=**en*) gefolgt wird; daran anschließend soll direkt ein Substantiv (POS=N) stehen, auf das ohne Distanzbeschränkung eine beliebige Realisierung des Lemmas *sein* (Lemma=*sein*) folgt."

[27] Vgl. z.B. Stonebraker et al. (2010), Lin/Dyer (2010) und Dean/Ghemawat (2004).

Unabhängig vom linguistischen Wert dieser Abfrage gewährt sie einen guten Einblick in die potenzielle Komplexität kombinierter Suchmuster. Insbesondere wird deutlich, dass sich auf diese Weise quasi-reguläre Suchmuster („regular expressions") sowie Platzhalterzeichen („wildcards") performant einbinden lassen.[28] Außerdem liefert die Aufteilung in einzelne Suchmuster einen gleichsam natürlichen Ansatz zur Parallelisierung. Unser in die Datenbank eingebettetes Framework kombiniert stets genau zwei einzelne Suchmuster zu einem Map-Prozess; diese Vorgabe orientiert sich an den in Kapitel 2.3.2 gewonnenen Erkenntnissen. Konkret bedeutet dies, dass Abfragen mit mehr als einem Join auf unserem Referenzsystem nicht mehr optimal skalieren – und deshalb möglichst vermieden werden. Ein totaler Verzicht auf Joins – d.h. die Zuweisung nur je eines Suchattributs zu einem Map-Prozess – würde andererseits das zu verarbeitende Datenvolumen unverhältnismäßig vergrößern und die Abfragezeiten in der Folge ebenfalls in die Höhe treiben.

Ebenfalls bereits in Kapitel 2.3 angesprochen haben wir die Notwendigkeit der Unterstützung von NOT-Abfragen. Ist beispielsweise als Bestandteil des Suchmusters festgelegt, dass an einer bestimmten Position im Korpus *kein* Substantiv erscheinen darf, würde die naheliegende Übersetzung in die SQL-Bedingung `where co_morpho != 'N'` die Nutzung von Indizes ausschließen und einen langwierigen „full table scan" initiieren. Wir umgehen dieses Hindernis durch die Segmentierung von NOT-Bedingungen in einzelne Map-Prozesse sowie die Formulierung positiver Einschränkungen. Übertragen auf unser Beispiel bedeutet das: Wir formulieren die gegenteilige Variante `where co_morpho = 'N'`, die den Wortklassen-Index effektiv nutzt und ihre Ergebnisse wie üblich temporär speichert. Spätere Reduce-Prozesse müssen dann lediglich darauf achten, dass der eigentlich gewünschte Ausschluss in der Aggregrationsphase auch tatsächlich stattfindet – idealerweise auf einer dann bereits deutlich reduzierten Datenmenge.

Die bei der oben beschriebenen Segmentierung und Paarbildung eventuell übrig bleibenden Einzelmuster (hier: Lemma=*sein*) werden geradewegs in eine Datenbank-View gemappt. Sämtliche Map-Prozesse können parallel ausgeführt werden und speichern ihre Ausgabe (Satznummer, Position des jeweils

[28] Zur Optimierung der Abfragezeiten haben wir diverse Indextypen evaluiert (B-Tree, Bitmap, zusammengesetzt, funktional). Die Ergebnisse dieser Testläufe sowie die daraus folgenden Maßnahmen zur Performance-Optimierung übersteigen jedoch den Rahmen der vorliegenden Publikation und bleiben einer zukünftigen Veröffentlichung vorbehalten. Einen exemplarischen Einblick in die Problematik unterschiedlicher Ausführungspläne für zusammengesetzte und Einspalten-Indizes bietet Abbildung 11.

ersten Suchmusters, Position des jeweils zweiten Suchmusters) in einer temporären Tabelle. Für Map1 wird also folgendes SQL-Statement generiert: `insert into TB_MAP1 (co_sentenceid, co_vorne, co_hinten) select t1.co_sentenceid, t1.co_id, t2.co_id from TB_CONNEXOR_TOKEN t1, TB_CONNEXOR_MORPHO t2 where t1.co_sentenceid = t2.co_sentenceid and t1.co_id+1 = t2.co_id and t1.co_token='mit' and t2.co_morpho='A'`. Die beiden abgefragten Tabellen – bzw. deren Unterpartitionen – werden gegebenenfalls unter Nutzung vorhandener Indizes mit einem Hash Join verknüpft. Map2 und Map3 verfahren analog, Map4 legt mit `create view _REDUCE4 as select co_sentenceid, co_id from TB_CONNEXOR_LEMMA where co_lemma='sein'` die passende View an.

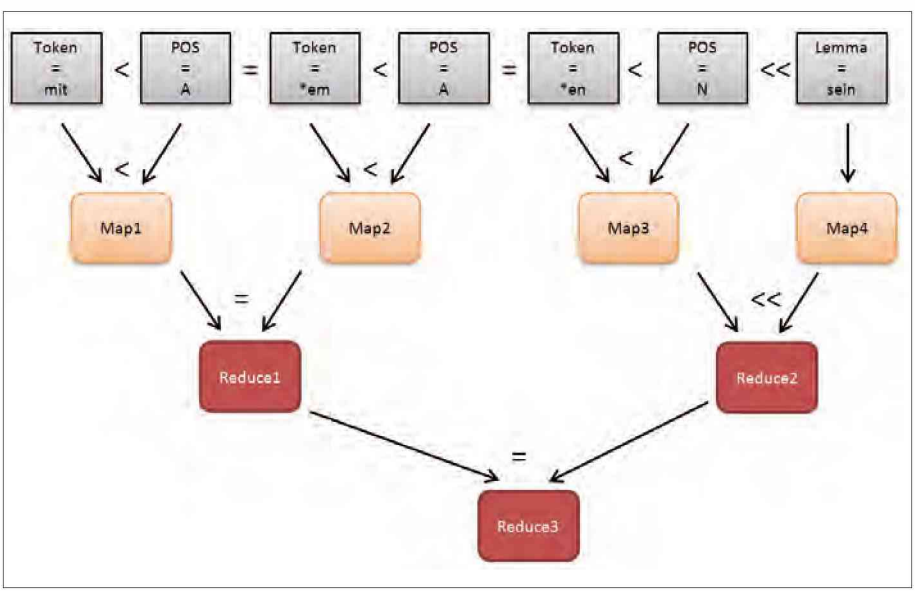



Abb. 10: MapReduce-Prozesse für eine Abfrage mit sieben Suchwerten

Ein Kerngedanke unseres Retrievalansatzes ist, dass die Parallelisierung nicht allein durch die Aufteilung in Map-Prozesse gesteuert werden kann. Als weiteres wichtiges Hilfsmittel stehen sogenannte „optimizer hints" zur Verfügung, die unser Framework in Abhängigkeit von der aktuellen Datenbankauslastung automatisch in die SQL-Statements einfügt. Erfordert eine Abfrage beispielsweise nur zwei Map-Prozesse, werden die entsprechenden SQL-Befehle mit

einem höheren Parallelitätsgrad ausgeführt als bei einer größeren Map-Anzahl. Übersteigt die Map-Anzahl die maximale CPU-/Thread-Grenze des Systems, kann der Parallelitätsgrad gezielt reduziert werden, um blockierende Sessions zu verhindern. Darüber hinaus evaluieren wir verschiedene regelbasierte Ansätze für die optimale Segmentierung komplexer Abfragen in Map-Prozesse, basierend auf quantitativen sprachlichen Parametern (also z.B. der Frequenz des abzufragenden Token/Lemma/Wortklassenbezeichners in der gespeicherten Grundgesamtheit). All diese Entscheidungen finden vollständig transparent im Hintergrund statt, d.h. der Benutzer muss sich nicht um die Steuerung der Parallelität kümmern.

Da die anschließenden Reduce-Prozesse zur Vermeidung von Deadlocks darauf angewiesen sind, dass alle relevanten Map-Prozesse sicher rechtzeitig terminieren,[29] übernimmt ein Datenbank-Scheduler das Anstoßen der SQL-Statements in der Map-Phase. Ist ein Map-Prozess beendet, so schreibt er eine entsprechende Meldung in eine Job-Tabelle. Reduce1 startet beispielsweise erst nach Abschluss von Map1 und Map2 und aggregiert die passenden Ergebnisse mit `insert into TB_REDUCE1 (co_sentenceid, co_vorne, co_hinten) select t1.co_sentenceid, t1.co_vorne, t2.co_hinten from TB_MAP1 t1, TB_MAP2 t2 where t1.co_sentenceid=t2.co_sentenceid and t1.co_hinten = t2.co_vorne`.[30] Die Spalte `co_vorne` nimmt dabei – wie schon vorher bei den Map-Prozessen – die Position des ersten Suchmusters auf, `co_hinten` die Position des letzten Suchmusters. Auf diese Weise lassen sich beliebig große Minimal- oder Maximalabstände zwischen den einzelnen Trefferstellen kodieren. Für die finale Ergebnistabelle, in der die Ausgabe von Reduce3 abgelegt wird, genügt dann die Speicherung der Satznummern: `insert into TB_REDUCE3 (co_sentenceid) select t1.co_sentenceid from TB_REDUCE1 t1, TB_REDUCE2 t2 where t1.co_sentenceid=t2.co_sentenceid and t1.co_hinten = t2.co_vorne`.

[29] Eine Ausnahme bildet naheliegenderweise ein Abfrageszenario mit nur einem oder zwei einzelnen Suchmustern. In diesem Fall liefert MAP1 bereits die endgültigen Ergebnisse, weitere Reduce-Schritte können entfallen.

[30] Bildet ein Map-Prozess eine NOT-Abfrage ab, muss an dieser Stelle – wie oben angedeutet – der Gleichheits- durch den Ungleichheitsoperator ersetzt werden.

### 2.5.2 Abfragen auf Textebene

Die bislang präsentierten Abfragen beschränkten sich auf Token-, Lemma- und Wortklassensuchmuster, die sich allesamt auf die Einzelwortebene beziehen (mit individuell möglichen Positionsabweichungen, weil verschiedene Parser in Einzelfällen eine unterschiedliche Tokenisierung liefern können). Komplettiert wird eine für die weitere statistische Auswertung aussagekräftige Korpusrecherche jedoch durch die Einbeziehung höherer Repräsentationsebenen, also bespielsweise Satz, Text oder Korpus. Unser Datenmodell verankert textspezifische Metadaten wie Datum, Domäne, Staat oder Register in der Tabelle TB_TEXT, die ihrerseits via TB_CONNEXOR_SENTENCE mit den Einzelwort-Tabellen verbunden ist. Soll ein Einzelwort-basierter Map-Prozess um entsprechende Einschränkungen erweitert werden, wäre also ein mehrstufiger Join notwendig, der selbst bei Nutzung von Spalten-Indizes vergleichsweise rechenintensiv bleibt. Aus diesem Grund lagern wir die Abfrage von Text-Metadaten in eigenständige Map-und Reduce-Prozesse aus, die ausschließlich auf der Satzebene operieren und parallel zu den Einzelwort-basierten Abfragephasen ablaufen.

```
create or replace function DomaeneConnexor (v_sentenceid in
    number) return number deterministic as

  v_wert number(1);

begin

  select co_domain into v_wert from TB_TEXT where co_textid in
    (select co_textid from TB_CONNEXOR_SENTENCE where
    co_sentenceid=v_sentenceid);

  return v_wert;

exception

  when others then return null;

end DomaeneConnexor;

create index idx_connexor_domaenesen on TB_CONNEXOR_SENTENCE
    (DomaeneConnexor(co_sentenceid), co_sentenceid) parallel 4
    nologging;

insert into TB_META_REDUCE1 select co_sentenceid from
    TB_CONNEXOR_SENTENCE where DomaeneConnexor (co_sentenceid) =
    2; -- Wert für Kultur/Unterhaltung
```

Listing 6: Anlegen und Nutzung des Domänen-Index

Listing 6 demonstriert das Anlegen und Abfragen der textbasierten Metadaten am Beispiel der Domänen-Klassifizierung. Da die Einzelwort-basierte Suchphase Satznummern zurückliefert, bietet sich das Anlegen eines zusammengesetzten funktionalen Index auf der Satztabelle an, der neben der Satznummer auch die Domänenangabe aufnimmt. Die entsprechende „Create Index"-Anweisung verwendet hierfür die in der Datenbank hinterlegte Funktion „DomaeneConnexor", die zu jeder Satznummer den zugehörigen numerischen Domänenwert ermittelt. Die initiale Indexerstellung benötigt auf unserem System bei einem Parallelitätsgrad von 4 zwar annähernd 30 Stunden, beschleunigt die spätere Abfrage jedoch substanziell.

Die beiden Ausführungspläne in Abbildung 11 verdeutlichen den signifikanten Vorteil eines zusammengesetzten Index für die Meta-Reduzierung. Die obere Abfrage nach sämtlichen Sätzen aus der Domäne 'Kultur/Unterhaltung' operiert ausschließlich auf dem zusammengesetzten Index IDX_CONNEXOR_DOMAENESEN, umgeht dadurch den „by index rowid"-Lookup in der Satztabelle und kopiert die insgesamt 103 480 892 Satznummern innerhalb von ca. 10 Sekunden in die temporäre Reduce-Tabelle (Abfragekosten: 3 823). Die untere Abfrage verwendet aufgrund des expliziten Index-Hints einen einspaltigen funktionalen Index (ohne Satznummern) und terminiert nach ca. 5 Minuten (Abfragekosten 185988).

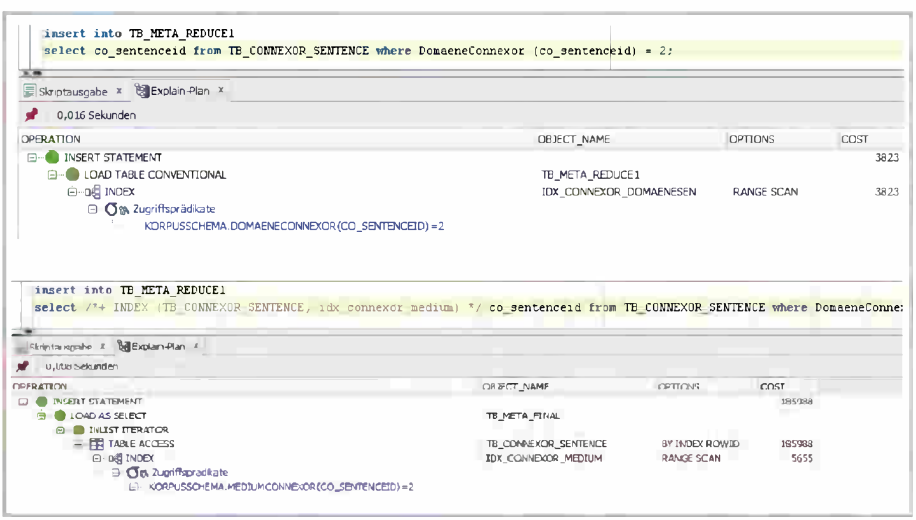



Abb. 11: Metadatenabfrage mit zusammengesetztem Index (oben) bzw. Einspalten-Index (unten)

Die Suche anhand textspezifischer Metadaten liefert, ebenso wie die Einzelwort-basierte Suchphase in 2.5.1, eine Liste passender Satznummern zurück. Nach Abschluss beider Phasen müssen deshalb nur noch die identischen Satznummern aus den TB_REDUCE- und TB_META_REDUCE-Tabellen ermittelt und in einem finalen Schritt zusammengezogen werden.

## 2.6 Die Abfrageoberfläche

Die *KoGra-DB*-Abfrageoberfläche ist als browsergestützte, plattformunabhängige Client-Software konzipiert, deren dynamische Webseiten passwortgeschützt im Intra-/Internet des IDS erreichbar sind. Die Generierung der Recherche- und Ergebnis-Webseiten erfolgt durch in der Datenbank gespeicherte Prozeduren („stored procedures"). Funktional entspricht dieser Ansatz der in der Software-Entwicklung klassischen Drei-Schichten-Architektur („three tier architecture") mit der Datenbank als Backend („data tier") für das persistente Speichern der Korpusinhalte, einem Datenbank-Web-Gateway als zwischengeschalteter Schicht für die Anwendungslogik („middle tier") und dem Web-Browser als Frontend für Benutzereingaben und Ergebnisrepräsentation („client tier"). Aufgrund der Nutzung von Stored Procedures existiert eine marginale technische Kopplung zwischen Backend und Logikschicht. Durch die strikte logische Trennung der jeweiligen Software- und Daten-Module wird diese Kopplung allerdings aufgefangen, so dass keinerlei Abhängigkeitsprobleme hinsichtlich der Pflege, Weiterentwicklung oder Skalierbarkeit einzelner Schichten zu erwarten sind.

Im Folgenden werden Funktion und Handhabung des Frontends anhand praktischer Einsatzszenarien vorgestellt.

### 2.6.1 Abfrage von Listen und Übersichten

Die *KoGra-DB*-Anwendungsoberfläche soll – neben der zielgerichteten Recherche – eine freie Navigation zwischen automatisch generierten Kontroll- und Übersichtslisten ermöglichen. Hierzu zählen beispielsweise variabel geordnete Listen der segmentierten linguistischen Einheiten (Wortformen, Lemmata, Sätze, Texte, Korpora u.A.) sowie deren Frequenz oder Länge ebenso wie quantitative Aussagen über Umfang und Distribution von Meta-Merkmalen.

Abbildung 12 demonstriert die Navigation innerhalb der bereits im Zusammenang mit der Modellierung der *KoGra-DB*-Tabellenstrukturen angesprochenen Tokenliste. Länge und Frequenz der insgesamt mehr als 20 Millionen unterschiedlichen Connexor-Wortformen liegen in einer mehrfach indizierten Lookup-Tabelle vor und können per Web-Formular als zusätzliche Ordnungskriterien neben der alphabetischen Reihenfolge ausgewählt werden. Satzzeichen werden hier übrigens wie Textwörter behandelt und erscheinen deshalb ebenfalls in der Tokenliste – Komma und Punkt bei Anordnung nach Frequenz erwartungsgemäß sogar auf den beiden vordersten Plätzen.[31]

**Tokenliste**

Tagger Connexor sortiert nach Frequenz

Alphabet
Frequenz
Länge

DEREKO-Release vom 29.03.2011: [illegible]xte: 255.755.622 Connexor-Sätze: 4.500.734.459 Connexor-Token.

[<<] 0-999/20278054 [>>]

| Frequenz | Länge | Token |
|---|---|---|
| 226992454 | 1 | . |
| 223566130 | 1 | , |
| 117112521 | 3 | der |
| 106554248 | 3 | die |
| 84071263 | 3 | und |
| 69883084 | 1 | in |
| 64942328 | 2 | in |
| 63565064 | 6 | " |
| 48189733 | 2 | zu |
| 43633702 | 3 | den |
| 34088396 | 3 | von |
| 32425256 | 3 | mit |
| 30790500 | 3 | das |
| 30763212 | 1 | - |
| 28066584 | 1 | I |
| 27615684 | 1 | – |
| 27429684 | 4 | sich |
| 26693133 | 3 | ist |
| 26297939 | 3 | auf |
| 26237513 | 3 | für |

**Abb. 12: Variabel geordnete Connexor-Tokenliste**

[31] Die Frequenzangaben weichen von den Zahlen in Kapitel 2.3.2 ab, da unser Gesamtkorpus umfangreicher als das dort verwendete 4-Milliarden-Korpus ist.

Die in Abbildung 13 dargestellte Abfrage von Umfang und Verteilung der verschiedenen textspezifischen Metadaten verzichtet dagegen angesichts der Vielzahl spezifizierbarer Kombinationen auf die Nutzung vorab berechneter statischer Daten. Meta-Kategorien wie Medium, Register, Domäne, Ort oder Jahr lassen sich stattdessen flexibel in Drop-Down-Listen auswählen und initiieren zur Laufzeit zielgerichtete Abfragen über das gesamte Korpusinventar; alternativ lassen sich die Abfragen auch auf einzelne Teilkopora beschränken. Innerhalb weniger Sekunden ermittelt das System die Anzahl der passenden Texte, Sätze oder Wörter (wobei die Abbildung nochmals verdeutlicht, dass hier Tagger-spezifische Tokenisierungen verwaltet werden, was z.B. die hohen Frequenzen für „m“ oder „i“ erklärt).

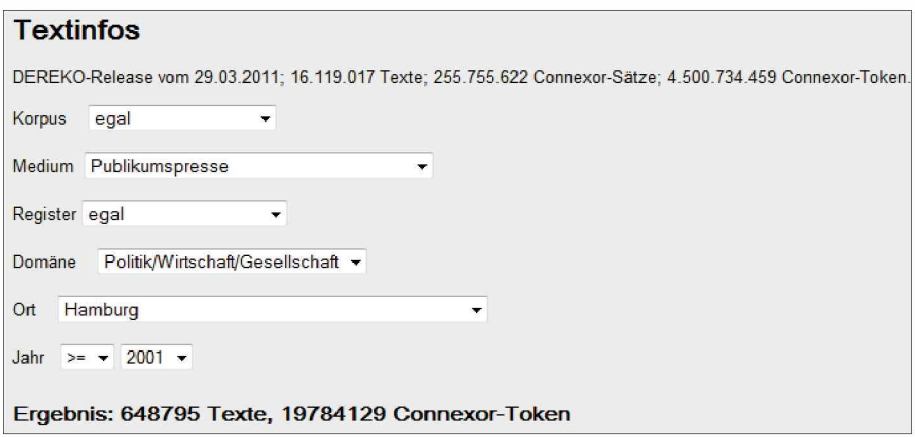
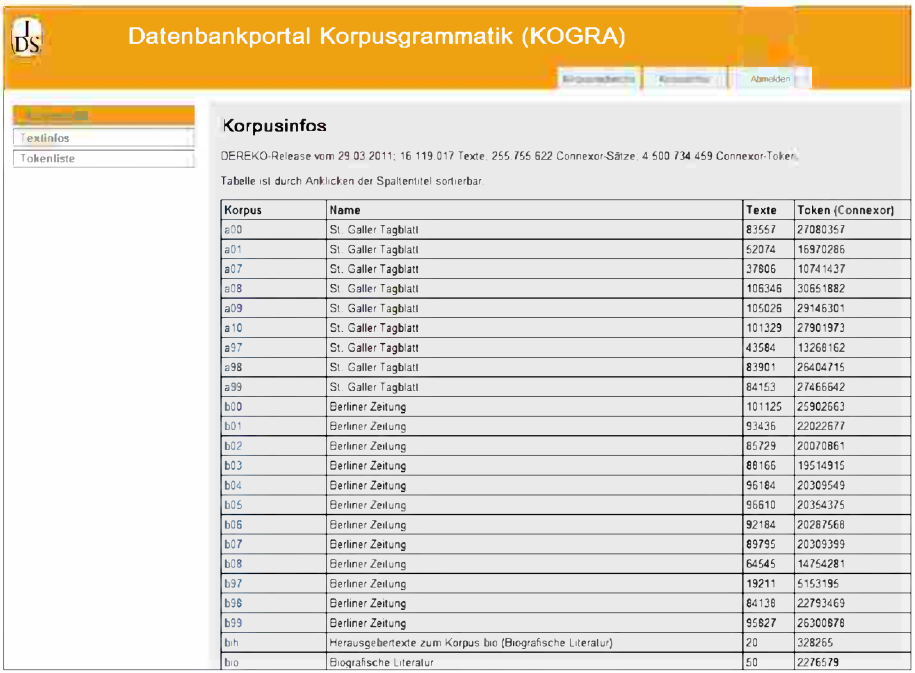



Abb. 13: Ad-hoc-Berechnung der Verteilung von Meta-Merkmalen

Oberhalb der Textebene kann sich der *KoGra*-Nutzer quantitative und Meta-Daten zu einzelnen Korpora bzw. Teilkopora anzeigen lassen; Abbildung 14 zeigt hierfür den Ausgangspunkt. Abgefragt werden korpusspezifische Look-up-Tabellen, die neben der Korpussigle den Namen des Korpus sowie die Anzahl der enthaltenen Texte bzw. Wortformen beinhalten. Die daraus generierten Übersichten lassen sich zur Abfragezeit flexibel nach jedem verfügbaren Spaltentyp anordnen.




| Korpus | Name | Texte | Token (Connexor) |
|---|---|---|---|
| a00 | St. Galler Tagblatt | 83557 | 27080357 |
| a01 | St. Galler Tagblatt | 52074 | 16970286 |
| a07 | St. Galler Tagblatt | 37806 | 10741437 |
| a08 | St. Galler Tagblatt | 106346 | 30651882 |
| a09 | St. Galler Tagblatt | 105026 | 29146301 |
| a10 | St. Galler Tagblatt | 101329 | 27901973 |
| a97 | St. Galler Tagblatt | 43584 | 13268162 |
| a98 | St. Galler Tagblatt | 83901 | 26404715 |
| a99 | St. Galler Tagblatt | 84153 | 27466642 |
| b00 | Berliner Zeitung | 101125 | 25902663 |
| b01 | Berliner Zeitung | 93436 | 22022677 |
| b02 | Berliner Zeitung | 85729 | 20070861 |
| b03 | Berliner Zeitung | 88166 | 19514915 |
| b04 | Berliner Zeitung | 96184 | 20309549 |
| b05 | Berliner Zeitung | 96610 | 20354375 |
| b06 | Berliner Zeitung | 92184 | 20287568 |
| b07 | Berliner Zeitung | 89795 | 20309399 |
| b08 | Berliner Zeitung | 64545 | 14764281 |
| b97 | Berliner Zeitung | 19211 | 5153195 |
| b98 | Berliner Zeitung | 84138 | 22793469 |
| b99 | Berliner Zeitung | 95827 | 26300878 |
| bih | Herausgebertexte zum Korpus bio (Biografische Literatur) | 20 | 328265 |
| bio | Biografische Literatur | 50 | 2276579 |

**Abb. 14: Übersicht der (Teil-)Korpora**

Durch das Anklicken einer Korpussigle lassen sich zusätzliche Angaben zum jeweiligen Einzelkorpus einblenden; Abbildung 15 illustriert die entsprechende Funktionalität. Jeweils 1 000 Texte eines Korpus werden dort nach Text-ID geordnet aufgelistet, in weiteren Spalten finden sich der Texttitel sowie textspezifische Metadaten zu Medium, Register, Domäne, Ort und Jahr.[32] Per Mausklick kann in dieser Liste vor- und zurückgeblättert oder auch das Ordnungskriterium manipuliert werden.

Ein Klick auf die Text-ID schließlich öffnet ein Unterfenster mit den eigentlichen Textdaten. Die Motivation hierfür begründet sich durch die Praxiserfahrung, dass es zur Beurteilung der Relevanz einer Textsammlung für die Beantwortung einer grammatischen Fragestellung gelegentlich sinnvoll ist, sich vorab einen unmittelbaren Überblick über die Inhalte zu verschaffen. Aber auch die manuelle Zuordnung von deskriptiven Meta-Kategorien zu Einzeltexten bzw. die nachträgliche Validierung automatisierter Text-Klassifizierungen profitieren nicht selten von einem gezielten Blick in die Primärdaten.

[32] Diese Auswahl orientiert sich an den momentanen Projektbedürfnissen und lässt sich bei Bedarf ohne nennenswerten Programmieraufwand erweitern oder reduzieren.

**Korpusinfos brz05**

DEREKO-Release vom 29.03.2011: 16.119.017 Texte; 255.755.622 Connexor-Sätze; 4.500.734.459 Connexor-Token.

Tabelle ist durch Anklicken der Spaltentitel sortierbar

[<<] 1-1000/49712 [>>]

| Text-ID | Titel | Medium | Register | Domäne | Heimatort | Jahr |
|---|---|---|---|---|---|---|
| BRZ05/DEZ.00001 | Braunschweiger Zeitung, 06.12.2005: Eröffnung für 14 Januar geplant | 1 | 1 | 2 | Braunschweig | 2005 |
| BRZ05/DEZ.00003 | Braunschweiger Zeitung, 06.12.2005: Friedenslicht aus Bethlehem trifft am Sonntag ein | 1 | 1 | 4 | Braunschweig | 2005 |
| BRZ05/DEZ.00004 | Braunschweiger Zeitung, 06.12.2005: Preis für die Volkshochschule | 1 | 1 | 4 | Braunschweig | 2005 |
| BRZ05/DEZ.00007 | Braunschweiger Zeitung, 06.12.2005: Behörden wollen Flughafen-Plan durchziehen | 1 | 1 | 4 | Braunschweig | 2005 |
| BRZ05/DEZ.00010 | Braunschweiger Zeitung, 06.12.2005: | 1 | 1 | 2 | Braunschweig | 2005 |

Abb. 15: Einzelkorpus-Informationen

## 2.6.2 Kombinierte Recherche

Unter der Bezeichnung „Kombinierte Recherche" bietet das Frontend eine zentrale Anlaufstelle für die gezielte musterbasierte Suche nach Belegsätzen unter Verwendung sämtlicher verfügbarer Beschreibungsebenen bzw. Metadaten. Ein dynamisches Web-Formular unterstützt die syntaktisch korrekte Eingabe von Suchanfragen, ohne dass vorab eine spezielle Abfragesprache gelernt werden müsste.[33] Eingegebene bzw. angekreuzte Suchkriterien werden, wie in Kapitel 2.5.1 beschrieben, in SQL-basierte Map-Prozesse übersetzt, lose gekoppelt und soweit möglich synchron verarbeitet.

Abbildung 16 zeigt die passend ausgefüllte Suchmaske für unsere Beispielabfrage aus Kapitel 2.5.1: „Finde alle Sätze, die das Token *mit* enthalten, das unmittelbar von einem auf *em* endenden Adjektiv gefolgt wird, welches seinerseits unmittelbar von einem Adjektiv mit der Endung *en* gefolgt wird; daran anschließend soll direkt ein Substantiv stehen, auf das ohne Distanzbeschränkung eine beliebige Realisierung des Lemmas *sein* folgt." Die Werte für Token und Lemma werden stets frei eingegeben, bei der Auswahl der Wortklassenbezeichner hilft eine inkrementelle Drop-Down-Liste.

[33] Leistungsfähigkeit und eventuelle Unzulänglichkeiten dieser benutzerfreundlichen Herangehensweise sollen im späteren Projektverlauf gezielt evaluiert werden, gegebenenfalls als Vorstufe zur Entwicklung einer rein textbasierten Abfragesprache.

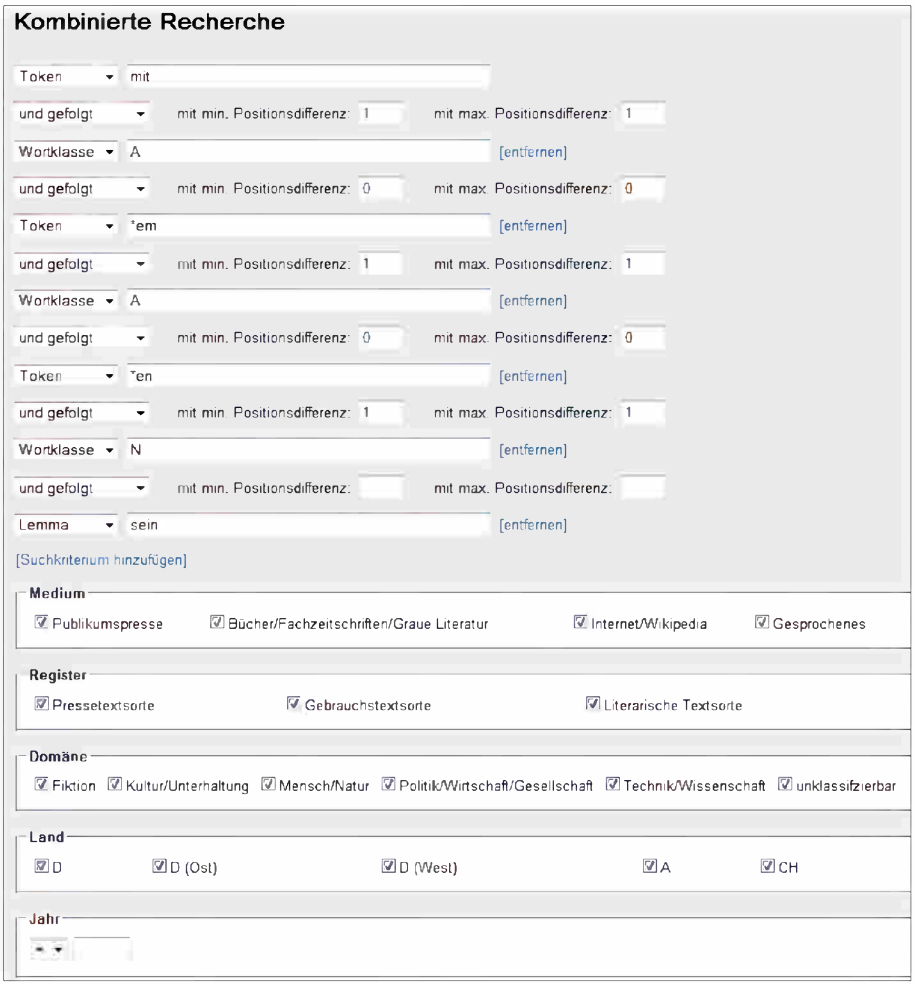



**Abb. 16: Beispielabfrage im Suchformular**

Minimale und maximale Abstände zwischen den einzelnen Suchkriterien können explizit angegeben werden. Bleiben die entsprechenden Felder leer (beispielsweise vor dem Kriterium „Lemma = *sein*"), so wird als Maximalabstand die Spanne bis zum Ende des jeweiligen Satzes angenommen. Möglich sind an dieser Stelle im Übrigen auch negative Werte: Ein Minimalabstand von „-5" legt fest, dass das folgende Suchkriterium mindestens fünf Positionen weiter vorne lokalisiert werden soll. Wird als Minimal- und Maximalwert die Null eingetragen, so impliziert dies Positionsidentität – auf diese Weise lassen sich Suchmuster wie „Adjektiv auf **em*" ausdrücken.

Unsere Beispielabfrage verwendet zwischen den einzelnen Suchkriterien ausschließlich „und gefolgt"-Verknüpfungen. Die Suchmaske bietet darüber hinaus für die Formulierung von NOT-Aussagen die Verknüpfungsbedingung „und nicht gefolgt". Bei der Eingabe von Abstandswerten für komplexe Suchanfragen gilt es in diesem Zusammenhang zu beachten, dass ein NOT-Kriterium keine Auswirkungen auf die Positionsangaben von möglicherweise folgenden Suchkriterien hat. Anders ausgedrückt: Eingebettete NOT-Kriterien beeinflussen nicht die für die Abfrage wichtige Differenz der Abstände von voranstehenden zu nachfolgenden Suchkriterien!

11:33:31 Suchprozess läuft...

| Prozess | Startzeit | Endzeit |
|---|---|---|
| map1-1 | 08.09.2011 11:30:41 | |
| map1-2 | 08.09.2011 11:30:41 | |
| map1-3 | 08.09.2011 11:30:41 | |
| map1-4 | 08.09.2011 11:30:41 | |
| map1-5 | 08.09.2011 11:30:41 | |
| map1-6 | 08.09.2011 11:30:41 | |
| map2-1 | 08.09.2011 11:30:41 | 08.09.2011 11:33:23 |
| map2-2 | 08.09.2011 11:30:41 | 08.09.2011 11:32:42 |
| map2-3 | 08.09.2011 11:30:41 | 08.09.2011 11:32:31 |
| map2-4 | 08.09.2011 11:30:41 | 08.09.2011 11:32:52 |
| map2-5 | 08.09.2011 11:30:41 | 08.09.2011 11:33:02 |
| map2-6 | 08.09.2011 11:30:41 | |
| map3-1 | 08.09.2011 11:30:41 | |
| map3-2 | 08.09.2011 11:30:41 | |
| map3-3 | 08.09.2011 11:30:41 | |
| map3-4 | 08.09.2011 11:30:41 | |
| map3-5 | 08.09.2011 11:30:41 | |
| map3-6 | 08.09.2011 11:30:41 | |
| map4 | 08.09.2011 11:30:41 | 08.09.2011 11:30:41 |

**Abb. 17: Statusmeldungen zur Beispielabfrage**

Unmittelbar nach Beginn des Suchprozesses werden die Namen und Startzeiten sämtlicher initiierten Map- und Reduce-Prozesse in die Suchseite eingeblendet (siehe Abbildung 17). Durch die in Kapitel 2.4 beschriebene physikalische Partitionierung der relevanten Datenbanktabellen können einzelne Map-Prozesse parallel in unterschiedlichen Korpusräumen suchen: Die Map-

Prozesse 1 bis 3 werden im vorliegenden Fall in jeweils sechs partitionsspezifische Einzelabfragen aufgeteilt („map1-1" bis „map1-6" usw.), Map-Prozess 4 erstellt mangels Join-Notwendigkeit einen einfachen Datenbankview. Start- und Endzeiten der an die Map-Phase anschließenden Reduce-Prozesse werden zur besseren Orientierung des Benutzers über den Stand der Recherche ebenfalls unmittelbar eingeblendet; Gleiches gilt für die Laufzeiten eventueller Map-/Reduce-Prozesse zur Einschränkung des Suchraums unter Heranziehung der textspezifischen Metadaten.

Die obige Abfrage über den gesamten verfügbaren Suchraum – also ca. 250 Millionen Sätze – findet insgesamt 1 597 Treffer; hier eine Auswahl.[34]

[1] *Wenn Sie mit angeklebtem weißen Bart und roter Zipfelmütze durch eine deutsche Stadt laufen , grüßt Sie – zumindest, wenn Weihnachtszeit ist – jeder, der eine Uniform anhat.*

[2] *Für den Club wiegt der Ausfall schwer , denn Golke ist ein Spieler mit großem taktischen Verständnis, der auch ohne Traineranweisung erkennt , wo er gerade gebraucht wird und wo Lücken zu schließen sind.*

[3] *Zu m " Anwerbe-Programm " gehören auch Maßnahmen, die mit gehörigem finanziellen Aufwand verbunden sind.*

[4] *Die Eigentümer hätten vor Mietabschlägen bei Wohnungen mit schlechtem energetischen Chychla verweist auf den Heizspiegel seines Verbandes , mit dem Mieter den Energieverbrauch ihrer Wohnung ermitteln können : wenden ihn bei unseren Mieterberatungen Sonnemann zitiert zudem ein Urteil des Landgerichts Hamburg vo m 11. September.*

Als http-basierte Web-Anwendung überträgt die Suchseite sämtliche spezifizierten Suchkriterien als dynamische Parameter per http-POST an den Datenbankserver. Ein zusätzlicher Stoppwert-Parameter gibt darüber Auskunft, ob die Recherche abgeschlossen oder noch auf dem Server aktiv ist, in Abhängigkeit davon werden entweder die ermittelten Ergebnisse eingeblendet oder – innerhalb eines angemessenen Zeitintervalls – das Suchformular neu geladen und die Statusangaben zu den laufenden Prozessen aktualisiert. Das Nassi-Shneiderman-Diagramm in Abbildung 18 präzisiert die dahinter liegende Programmlogik.

[34] Die Segmentierungen beruhen wiederum auf der Connexor-Tokenisierung; deshalb z.B. „vo m" statt „vom" in Treffer [4].

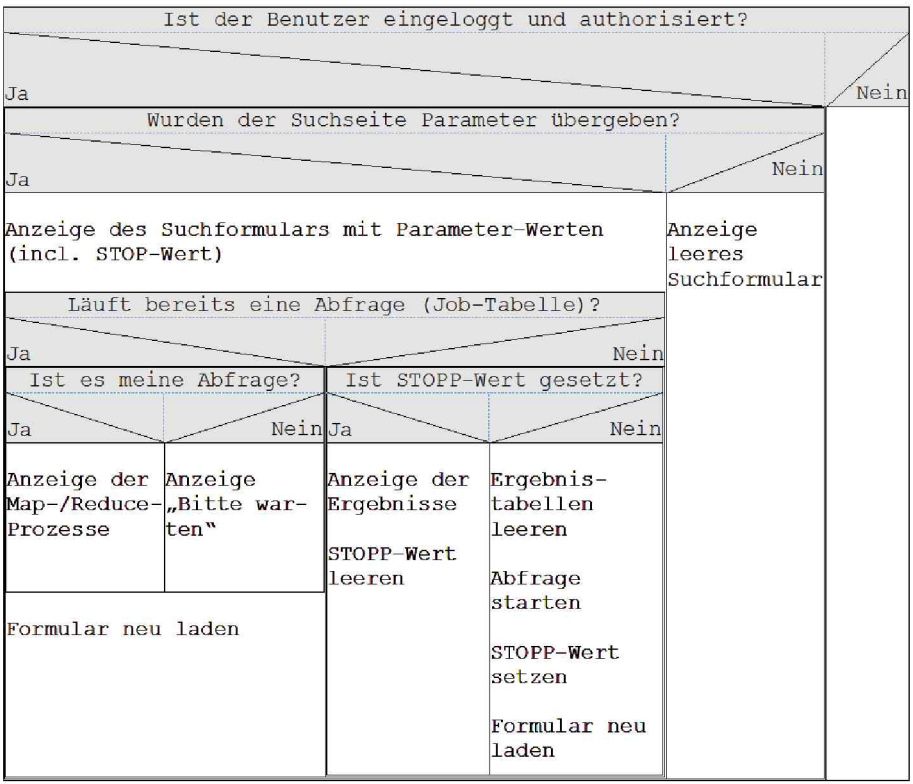



**Abb. 18: Programmlogik der Suchseite**

Um die Skalierfähigkeit unseres Abfrage-Frameworks zu überprüfen, lohnt sich nach Abschluss der Recherche ein Blick in die Administrations-Werkzeuge des Datenbankmanagementsystems. Abbildung 19 illustriert die Serverauslastung während der Beispielabfrage: Offensichtlich verteilen sich die Map- und Reduce-Prozesse wie gewünscht auf parallele Sessions, die allerdings aufgrund der vergleichsweise geringen Anzahl an CPU-Kernen via Hyperthreading abgearbeitet werden müssen. Eine höhere CPU-Zahl würde an dieser Stelle die anfallende Last optimaler verteilen und die Abfragezeit signifikant reduzieren.[35]

[35] Zur Begründung dieser plausiblen Annahme sollen in Kürze entsprechende Versuchsreihen auf einem Mehrprozessorsystem initiiert werden, erste Ergebnisse präsentiert Schneider (2012).

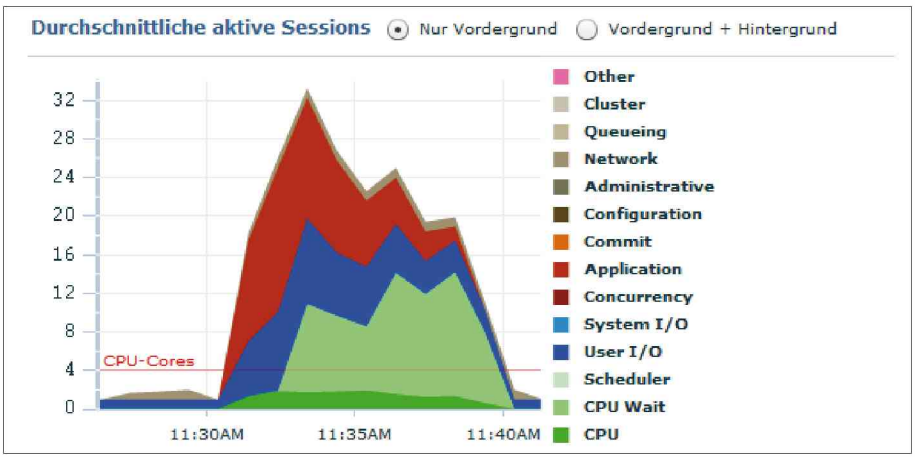



Abb. 19: Serverauslastung während der Beispielabfrage

Um eine effektive Nachnutzung von Suchergebnissen zu ermöglichen, speichert das Recherche-Frontend auf Wunsch die relevanten Abfragedaten (Suchmuster, Suchraum, Trefferzahl) dauerhaft in benutzerspezifischen Ergebnistabellen. Die Belege – auch in annotierter Form – für derart abgelegte Abfragen lassen sich anschließend jederzeit wieder aufrufen.

Gespeicherte Recherchen von Benutzer *roman*

| | Name | Datum | Typ | Abfrage | Suchraum | Treffer | Verteilung | |
|---|---|---|---|---|---|---|---|---|
| ☐ | Advanced Query | 08.09.2011 | C | token(mit) _u_morpho(A) _u_token(*em) _u_morpho(A) _u_token(*en) _u_morpho(N) _u_lemma(sein) | M(1,2,3,4)R(1,2,3)D(1,2,3,4,5,6) L(1,2,3,4,5)U(6,7,8,9,0,1) | 1697 | **MEDIUM:**Publikumspresse,1438/244245810:Bücher,130 /7133328:Internet,29/4183210:Gesprochenes,0/193126, **REGISTER:**Presse,1438/244245810:Gebrauch,32 /4593956:Literarisch,4/222176, **DOMAENE:**Fiktion,5/364223:Kultur,607/103480892:Mensch,45 /6165203:Politik,583/105369486:Technik,115 /17412207:unklassifizierbar,242/22983463, **LAND:**D,1172/17908921:DOst,2/187512:DWest,40/4005620:A,212 /43578759:CH,58/22027182, **JAHR:**1970,123/6420582:1970-79,0/164463:1980-89,28 /2868378:1990-99,619/90402568:2000-09,761/140528263:2010-,66 /15193467 | [Belege] [Annotierte Belege] [Erweiterte Suche] [R-Export] [löschen] |

Abb. 20: Gespeicherte Beispielabfrage mit Verteilungsangaben

Abbildung 20 zeigt die Benutzersicht auf gespeicherte Abfragen. Die Schaltfläche „Markierte Abfragen zusammenfassen" ermöglicht die Zusammenführung von Ergebnismengen, sofern der Suchraum der betreffenden Abfragen übereinstimmt. Auf diese Weise lässt sich dem Umstand begegnen, dass die derzeitige Abfragemaske kein Boole'sches „oder" (z.B. Suche nach Token *frug* oder *fragte*) und keine Optionalitätsspezifikation (z.B. Suche nach Artikeln und Substantiven, die entweder unmittelbar aufeinander folgen oder optional durch ein Adjektiv getrennt sind) anbietet: Die entsprechenden Abfragen können nacheinander ausgeführt und die gefundenen Treffer anschließend kombiniert werden.

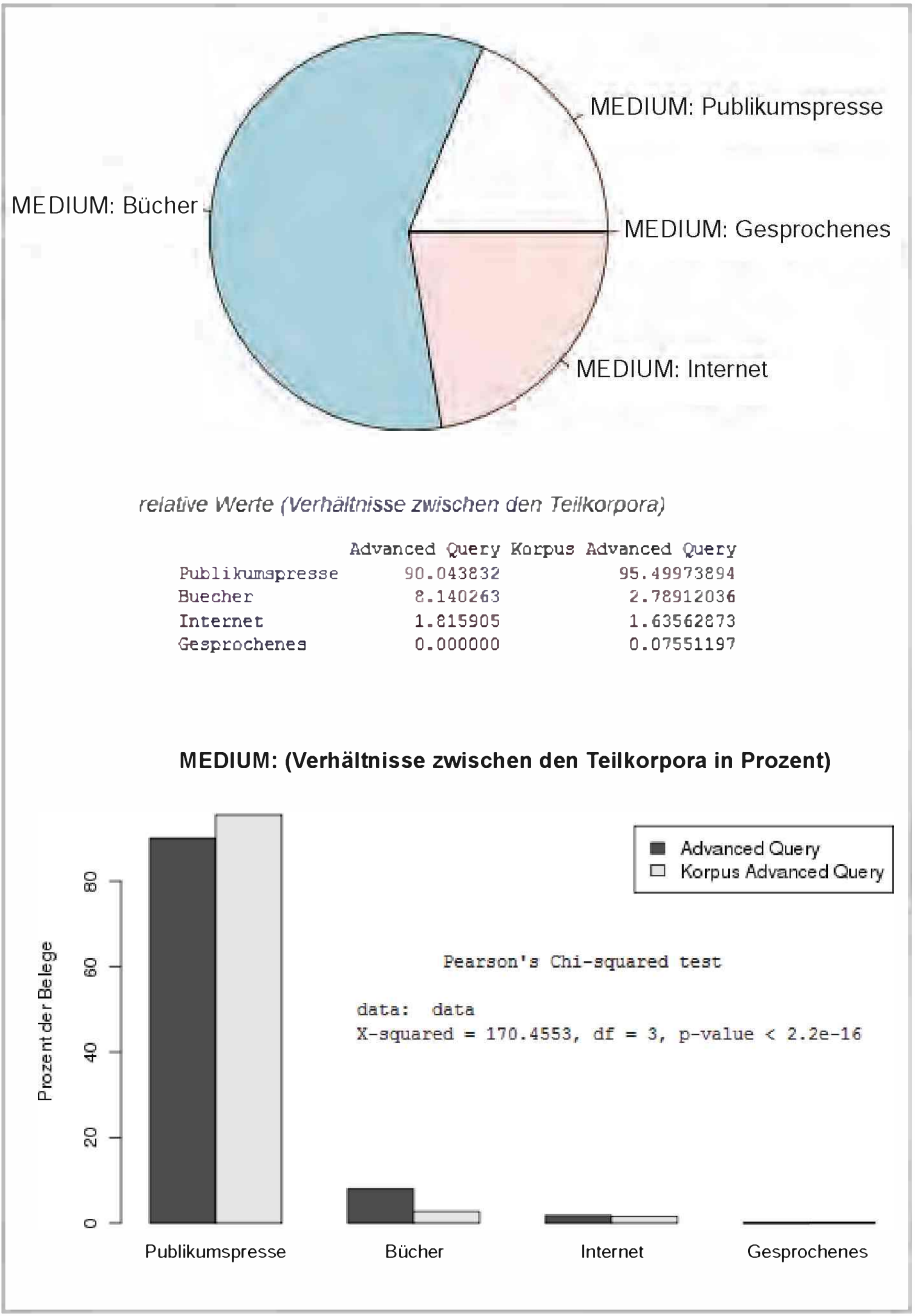



Abb. 21: Weiterverarbeitung der Verteilungsangaben in R

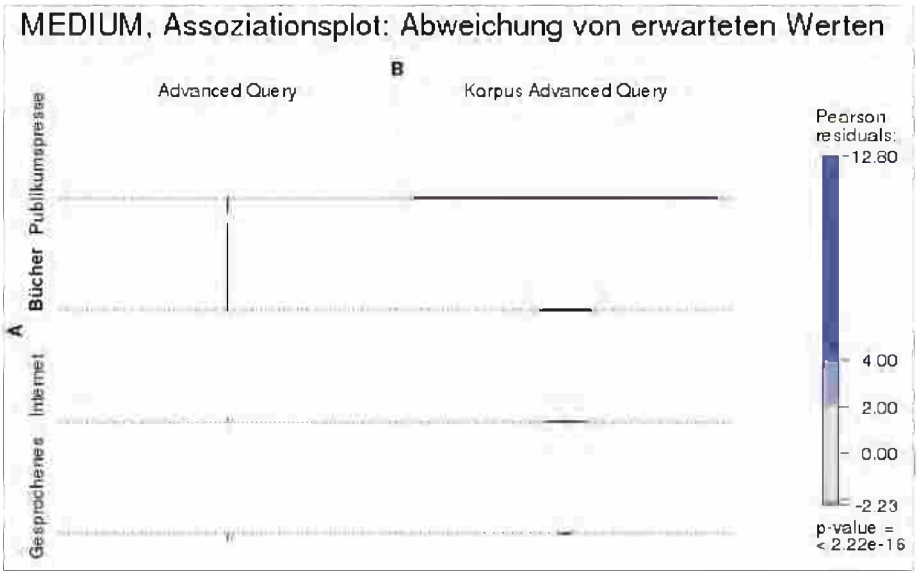



**Abb. 22: Weiterverarbeitung der Verteilungsangaben in R (Fortsetzung)**

Über die Option „Erweiterte Suche" lassen sich die Ergebnislisten der gespeicherten Abfragen unter Ausnutzung regulärer Suchausdrücke gezielt einschränken. Darüber hinaus können detaillierte Verteilungsangaben bezüglich Medium, Register, Domäne und Ort berechnet werden. Eine Export-Funktion für externe Werkzeuge zur statistischen Datenanalyse[36] (siehe Abbildungen 21 und 22 sowie die Beispiele und Visualisierungen in Kapitel 3.4) komplettiert den Funktionsumfang.

[36] Im vorliegenden Fall kommt das frei verfügbare „R" zum Einsatz; vgl. http://www.r-project.org.

# 3. Statistische Analysen in den Projektkorpora

Die Recherchen in den Projektkorpora sollen bis zu einem gewissen Maß vereinheitlicht werden: Ziel ist es, jeweils eine Reihe von immer gleichen statistischen Tests auf die Analyseresultate anzuwenden, um systematisch zu prüfen, ob die Verteilung der Treffer über die verschiedenen Teilkorpora in einem Bereich statistisch auffällig ist oder nicht. Dies dient dazu, ein Phänomen rasch einschätzen zu können: Handelt es sich um ein Randphänomen oder ist es in den Daten weit verbreitet? Ist es nur in bestimmten Textbereichen gebräuchlich? Aufgrund dieses Schnelltests können dann weitere Analysen vorgenommen werden.

## 3.1 Ausgangslage

Um ein systematisches Vorgehen bei den Recherchen in den Projektkorpora zu begünstigen, wurde das Recherchemodul (siehe Kapitel 2.1) um ein Modul der statistischen Analyse ergänzt, das automatisiert eine Reihe von Analysen durchführt. Diese Analysen geben bei der Recherche einen ersten Eindruck über die Verbreitung des Phänomens in den Daten und dienen als Basis für weitere Analysen.

Grundsätzlich wird bei diesen statistischen Analysen von zwei möglichen Rechercheszenarien ausgegangen:

1) Recherche nach einem bestimmten Phänomen (z.B. Verbzweitstellung in Nebensätzen)

2) Recherche nach zwei zusammenhängenden Phänomenen (z.B. Präteritum-Formen *fragte* vs. *frug*

Bei beiden Szenarien interessiert man sich dafür, wie hoch die Trefferzahl im Korpus ist und wie sich die Belege über das Korpus verteilen: Sind die Frequenzen in Bezug zur Korpusgröße eher hoch, mittel oder niedrig? Gibt es in einem oder mehreren der Teilkorpora, die über die Kriterien 'Medium', 'Region', 'Domäne' etc. (siehe Kapitel 1.4) definiert sind, eine auffällige Häufung der Treffer? Oder ist das Phänomen gleichmäßig über das Korpus verteilt? Es handelt sich also um Fragen der Varianz eines Phänomens in den Daten (vgl. Kapitel 1.2.2).

Beim zweiten Szenario, bei dem zwei Phänomene miteinander verglichen werden, interessiert zusätzlich noch das Verhältnis der Frequenzen zwischen den beiden Phänomenen. Ist das eine Phänomen durchgehend häufiger als das andere oder gibt es Teilkorpora, in denen das Verhältnis umgekehrt ist? Ausgangspunkt ist also ein Phänomen der Variation (vgl. Kapitel 1.2.1).

Diese Fragen sollen über einen „statistischen Schnelltest" bei jeder Recherche automatisch vom Korpusabfragesystem beantwortet werden. Die Beantwortung dieser Fragen gehört zum Standardrepertoire von Korpusanalysen (vgl. etwa Gries 2008b: 103ff., 153ff.; Oakes 1998; Albert/Koster 2002: 74ff.).

In diesem Kapitel wird dargestellt, welche statistischen Tests verwendet werden, um diesen „Schnelltest" durchzuführen, und wie sie technisch in R implementiert wurden.

## 3.2 Frequenzvergleiche zwischen Teilkorpora

Die Projektkorpora bestehen aus dem kleineren ausgewogenen Korpus und dem Gesamtkorpus. Beide lassen sich bezüglich verschiedener Metadaten in Teilkorpora unterteilen, so etwa medial oder regional (vgl. Kapitel 1.4). Die Korpusdatenbank (vgl. Kapitel 2) gibt bei einer Recherche für jedes dieser Teilkorpora die Trefferzahl und die jeweilige Gesamtanzahl der Wörter und Sätze in den Teilkorpora zurück. Damit lassen sich nun eine Reihe von Frequenzvergleichen durchführen:

**Fragestellung 1:**

Ist das Phänomen X gleichmäßig über die Teilkorpora verteilt oder gibt es in einem oder mehreren Teilkorpora unerwartet viele oder wenige Treffer?

**Fragestellung 2:**

Wie verhalten sich die Treffer der Phänomene X und Y zueinander? Ist der Unterschied generell oder in bestimmten Teilkorpora unerwartet hoch?

**Fragestellung 3:**

Wir kennen die Verteilung eines Phänomens X in einem Korpus A. Verteilt sich dieses Phänomen X in ähnlicher Weise auf die Teilkorpora von Korpus B oder gibt es in einem oder mehreren Teilkorpora unerwartet hohe Unterschiede?

Um Frequenzunterschiede zwischen Korpora zu messen, kann der nichtparametrische Pearsons Chi-Quadrat-Test verwendet werden (Manning/Schütze 2002: 169f.; Sheskin 2007: 619ff.). Kilgarriff (2001: 121ff.) zeigt, dass der Chi-Quadrat-Test gerade für den Vergleich von Frequenzen in Korpora gut geeignet ist.[1] Grundlage für den Chi-Quadrat-Test sind Kontingenztabellen, die die beobachteten und erwarteten Frequenzwerte für das Phänomen in den Teilkorpora enthalten (Sheskin 2007: 622ff.). Dabei wird geprüft, ob für die Grundgesamtheit auf der Basis der Zufallsauswahl die Nullhypothese $H_0$ gilt:

> **$H_0$:** Für jede Zelle der Kontingenztabelle gilt, dass sich die beobachteten nicht von den erwarteten Werten unterscheiden.

Entsprechend lautet die Alternativhypothese:

> **$H_1$:** Mindestens für eine Zelle der Kontingenztabelle unterscheidet sich der beobachtete vom erwarteten Wert.

Im Folgenden wird an den drei Fragestellungen gezeigt, wie die statistischen Berechnungen erfolgen. Es werden dabei fingierte Frequenzen verwendet, um die Berechnungen an einfachen Zahlenverhältnissen nachvollziehbar zu machen.

### 3.2.1 Verfahren Fragestellung 1 (X in den Teilkorpora)

Wir gehen von einem Phänomen X aus, das insgesamt 1 000 Mal im Korpus vorkommt. Nun wird für jede Kategorie, mit der das Korpus in Teilkorpora aufgeteilt werden kann, die Verteilung auf die Teilkorpora geprüft. Wir gehen davon aus, dass sich z.B. für die Kategorie 'Medium' eine Verteilung von je 100 Treffern in 'Publikumspresse', 'Bücher', 'Internet/Wikipedia' und 'Sonstige Printmedien' sowie von 700 Treffern in 'Gesprochenes' ergeben haben. Diese Werte sind in Tabelle 1 ersichtlich, wobei zusätzlich für jedes Teilkorpus die Gesamtgröße in Anzahl laufender Wortformen (Spalte „Total") und daraus abgeleitet die Anzahl laufender Wortformen, die nicht zu Phänomen X gehören (Spalte ¬X) eingetragen sind.

[1] Bei niedrigen Werten in der Kontingenztabelle (erwarteter Wert einer Zelle < 5, n < 20) sollte anstelle des Chi-Quadrat-Tests der exakte Test nach Fisher verwendet werden (Sheskin 2007: 631).

| Teilkorpus | X | ¬X | Total |
|---|---|---|---|
| Publikumspresse | 100 | 3 999 900 | 4 000 000 |
| Bücher | 100 | 3 999 900 | 4 000 000 |
| Internet/Wikipedia | 100 | 3 999 900 | 4 000 000 |
| Gesprochenes | 700 | 3 999 300 | 4 000 000 |
| Sonstige Printmedien | 100 | 3 999 900 | 4 000 000 |
| **Total** | **1 100** | **19 998 900** | **20 000 000** |

**Tab. 1: Beobachtete Werte von Phänomen X in den Teilkorpora**

Zusätzlich zur Tabelle mit den beobachteten Werten wird eine Tabelle mit den erwarteten Werten erstellt. Grundsätzlich könnte diese Tabelle beliebige Erwartungen an die Verteilung widerspiegeln. In unserem Fall gehen wir davon aus, dass die Treffer gleichmäßig auf die Teilkorpora verteilt sind. Die insgesamt 1100 Treffer müssten dann im Verhältnis zur jeweiligen Größe des Teilkorpus gleichmäßig auf die Teilkorpora verteilt sein, wie das in Tabelle 2 ersichtlich ist.[2] Da die Teilkorpora alle gleich groß sind, erwarten wir, dass die 1100 Treffer gleichmäßig auf die Teilkorpora verteilt sein müssten.

| Teilkorpus | X | ¬X | Total |
|---|---|---|---|
| Publikumspresse | 220 | 3 999 780 | 4 000 000 |
| Bücher | 220 | 3 999 780 | 4 000 000 |
| Internet/Wikipedia | 220 | 3 999 780 | 4 000 000 |
| Gesprochenes | 220 | 3 999 780 | 4 000 000 |
| Sonstige Printmedien | 220 | 3 999 780 | 4 000 000 |
| **Total** | **1 100** | **19 998 900** | **20 000 000** |

**Tab. 2: Erwartete Werte von Phänomen X in den Teilkorpora**

Der Chi-Quadrat-Wert ($\chi^2$) wird nun wie folgt berechnet, wobei *O* für die beobachteten und *E* für die erwarteten Werte stehen:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

[2] Der Wert für 'Publikumspresse' berechnet sich z.B. in einem Dreisatz wie folgt: 4 000 000 · 1 100 : 20 000 000 = 220.

Für jede Zelle der Kontingenztabelle wird also der beobachtete Wert (vgl. Tabelle 1) vom erwarteten Wert (vgl. Tabelle 2) subtrahiert, quadriert und durch den erwarteten Wert dividiert. Die Summe über alle Zellen ergibt $\chi^2$.

Jetzt kann geprüft werden, mit welcher Wahrscheinlichkeit p die Nullhypothese $H_0$, die eine zufällige Verteilung behauptet, verworfen werden kann. Dafür müssen die Freiheitsgrade df bestimmt werden, die sich wie folgt berechnen lassen: (Zeilenzahl – 1) · (Spaltenzahl – 1).[3] Für die oben dargestellte Kontingenztabelle beträgt dieser Wert also 4. Die Tabelle der kritischen Werte[4] für $\chi^2$ gibt dann Auskunft über die minimale Höhe, die $\chi^2$ für ein bestimmtes Signifikanzniveau haben muss, um $H_0$ verwerfen zu können.

Im Fall der oben genannten Werte ergibt $\chi^2$ den Wert 1309 und liegt damit über dem kritischen Wert von 18,467 für das Signifikanzniveau p = 0,001 und df = 4. Die Verteilung ist also mit einer Wahrscheinlichkeit von 99,9% nicht zufällig bzw. gleichmäßig. Normalerweise werden drei Signifikanzniveaus unterschieden: Wenn $p < 0,05$ spricht man von einem 'signifikanten' Unterschied. Wenn $p < 0,01$ ist von einem 'sehr signifikanten' und ab $p < 0,001$ von einem 'hoch signifikanten' Unterschied die Rede.

Die Höhe des Chi-Quadrat-Wertes sagt nichts über die Stärke der Assoziation zwischen Teilkorpus und Frequenz des Phänomens aus, da der Wert von der Größe der Kontingenztabelle abhängig ist. Deshalb wird Cramérs *V* zur Bereinigung des Wertes um die Größe der Kontingenztabelle verwendet:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

Dabei ist *N* die Gesamtzahl der beobachteten Werte und *k* die kleinere Zahl der Spalten oder Reihen der Kontingenztabelle. Im Falle der Berechnungen oben ergibt *V* den Wert 0,008. Da sich *V* immer zwischen 0 (keine Korrelation) und 1 (starke Korrelation) bewegt, muss die Korrelation im Beispiel oben als sehr schwach eingestuft werden.

Der Chi-Quadrat-Test sagt allerdings nur aus, dass die Treffer in irgendeinem oder mehreren Teilkorpora signifikant über oder unter den erwarteten Werten liegen, nicht aber, in welchen Teilkorpora das der Fall ist. Dazu kann nun ein paarweiser Vergleich durchgeführt werden (Sheskin 2007: 650): Es werden

[3] Die Total-Zeile und -Spalte wird jeweils nicht mitgezählt.

[4] Vgl. z.B. Manning/Schütze (2002: 609).

für alle Kombinationen Publikumspresse/Bücher, Publikumspresse/Internet, Publikumspresse/Gesprochenes, Bücher/Publikumspresse etc. Tests berechnet, so dass angegeben werden kann, welche Teilkorpora signifikant von den erwarteten Werten abweichen.

| | Publikums-presse | Bücher | Internet/ Wikipedia | Gesprochenes |
|---|---|---|---|---|
| **Bücher** | 1 | – | – | – |
| **Internet/Wikipedia** | 1 | 1 | – | – |
| **Gesprochenes** | < 2e–16 | < 2e–16 | <2e–16 | – |
| **Sonstige Printmedien** | 1 | 1 | 1 | <2e–16 |

Tab. 3: Paarweiser Vergleich, Werte für p (mit Bonferroni-Korrektur)

Tabelle 3 zeigt die Resultate des paarweisen Vergleichs und gibt für jede Kombination der Teilkorpora den Wert für p an. Wie man sieht, sind nur die Frequenzunterschiede zwischen dem Teilkorpus 'Gesprochenes' und den anderen hoch signifikant ($p < 2e\text{-}16$), die Unterschiede zwischen den anderen Teilkorpora sind nicht signifikant ($p = 1$).

Im Beispiel oben wurde das zu zählende Phänomen nicht weiter bestimmt: Es könnte sich um ein bestimmtes Wort, eine Konstruktion oder eine andere sprachliche Einheit handeln. Davon hängt aber ab, in welche Relation die Frequenz des Phänomens gesetzt werden soll: Sind die Summen in den Kontingenztabellen Frequenzen laufender Wortformen (wie im Beispiel oben) oder muss eine andere Einheit gewählt werden? Die Anzahl von Fällen von Verbzweitstellung in Nebensätzen hängt ja beispielsweise weniger von der Anzahl laufender Wortformen im Korpus ab, als von der Anzahl der Nebensätze im jeweiligen Korpus. Gibt es in einem Korpus nur sehr wenige Nebensätze, sind auch nicht viele Fälle von Verbzweitstellung zu erwarten, auch wenn das Korpus sehr viele laufende Wortformen umfasst.

Daher ist es sinnvoll, die zu zählenden Elemente flexibel zu halten. Der Einfachheit halber wird aber bei der statistischen Standardauswertung entweder die Anzahl laufender Wortformen oder die Anzahl an Sätzen berücksichtigt. Ist das bei einer bestimmten Fragestellung problematisch, kann dies manuell verändert werden.

### 3.2.2 Verfahren Fragestellung 2 (Verhältnis zwischen X und Y)

Um zwei Phänomene gegenüberzustellen, werden leicht modifzierte Kontingenztabellen erstellt (Tabellen 4 und 5).

| Teilkorpus | X | Y | Total |
|---|---|---|---|
| Publikumspresse | 100 | 200 | 300 |
| Bücher | 100 | 100 | 200 |
| Internet/Wikipedia | 100 | 100 | 200 |
| Gesprochenes | 700 | 100 | 800 |
| Sonstige Printmedien | 200 | 100 | 300 |
| **Total** | **1 200** | **600** | **1 800** |

Tab. 4: Beobachtete Werte für die Phänomene X und Y in den Teilkorpora

| Teilkorpus | X | Y | Total |
|---|---|---|---|
| Publikumspresse | 200 | 100 | 300 |
| Bücher | 133,3333 | 66,66667 | 200 |
| Internet/Wikipedia | 133,3333 | 66,66667 | 200 |
| Gesprochenes | 533,3333 | 266,66667 | 800 |
| Sonstige Printmedien | 200 | 100 | 300 |
| **Total** | **1 200** | **600** | **1 800** |

Tab. 5: Erwartete Werte für die Phänomene X und Y in den Teilkorpora

Auch hier gibt der Chi-Quadrat-Test nur darüber Auskunft, ob die Verteilung der Treffer für X und Y in Relation zu den jeweiligen Treffersummen von X und Y gleichmäßig verteilt sind. Bei den Messwerten in Tabelle 4 ergibt der Chi-Quadrat-Test eine hoch signifikante Abweichung von einer gleichmäßigen Verteilung ($p < 2{,}2e\text{-}16$, $\chi^2 = 356$, $df = 4$). Ist die Verteilung grundsätzlich ungleichmäßig, kann wieder mit paarweisen Vergleichen (X und Y in Publikumspresse und Bücher, Publikumspresse und Internet/Wikipedia etc.) ermittelt werden, in welchem Teilkorpus der Frequenzunterschied signifikant ist.

| | Publikums-presse | Bücher | Internet/ Wikipedia | Gesprochenes |
|---|---|---|---|---|
| **Bücher** | 0,0028 | – | – | – |
| **Internet/Wikipedia** | 0,0028 | 1,0000 | – | – |
| **Gesprochenes** | < 2e–16 | < 2e–16 | < 2e–16 | – |
| **Sonstige Printmedien** | 6,30e–15 | 0,0028 | 0,0028 | 3,00e–14 |

Tab. 6: Paarweiser Vergleich, Werte für p (mit Bonferroni-Korrektur)

Die Ergebnisse des paarweisen Vergleichs in Tabelle 6 zeigen, dass die Verteilung von X und Y im Teilkorpus 'Gesprochenenes' sich zu jedem anderen Teilkorpus hoch signifikant unterscheidet. Auch die Verteilungen von X und Y im Bücher-Korpus und dem Korpus der sonstigen Printmedien unterscheiden sich sehr stark. Andererseits ergibt sich zwischen den Korpora Bücher und Internet/Wikipedia überhaupt kein Unterschied (p = 1). Die anderen Vergleiche weisen mit p = 0,0028 ebenfalls (wenn auch nicht hoch) signifikante Unterschiede auf.

### 3.2.3 Verfahren Fragestellung 3 (Verhältnis der Verteilungen von X in Korpus A und B)

Bei Fragestellung 3 geht es darum, die Verteilung eines Phänomens X, das wir aber in zwei unterschiedlichen Korpora mit jeweils gleicher Untergliederung in Teilkorpora messen, zu vergleichen. Dieses Problem stellt sich typischerweise, wenn die Verteilung eines Phänomens im ausgewogenen Korpus mit der Verteilung im Gesamtkorpus verglichen werden soll. Wenn das ausgewogene Korpus eine zufällige Stichprobe aus dem Gesamtkorpus wäre, müsste die Verteilung der Frequenzen von X auf die Teilkorpora sehr ähnlich sein. Der Kontingenztabelle von Phänomen X im Gesamtkorpus wird eine Tabelle der erwarteten Werte gegenübergestellt, die auf der Verteilung des Phänomens X im ausgewogenen Korpus entspricht.

Gehen wir von zwei Korpora A und B aus. Die Treffer von X in Korpus A entspricht der Tabelle 1 oben. Korpus B sei nun insgesamt 10 mal größer als Korpus A und bezüglich seiner Teilkorpora gleich strukturiert. Nun seien die Treffer für X in Korpus B idealerweise ebenfalls jeweils das Zehnfache der Treffer in Korpus A, so dass die Kontingenztabelle 7 erstellt werden kann.

| Korpus B | X | ¬X | Total |
|---|---|---|---|
| Publikumspresse | 1 000 | 39 999 000 | 40 000 000 |
| Bücher | 1 000 | 39 999 000 | 40 000 000 |
| Internet/Wikipedia | 1 000 | 39 999 000 | 40 000 000 |
| Gesprochenes | 7 000 | 39 993 000 | 40 000 000 |
| Sonstige Printmedien | 1 000 | 39 999 000 | 40 000 000 |
| **Total** | **11 000** | **199 989 000** | **200 000 000** |

Tab. 7: Beobachtete Werte von Phänomen X in den Teilkorpora von Korpus B

Da wir die beobachteten Werte von X in Korpus A kennen (Tabelle 1), können wir nun daraus die für Korpus B erwarteten Werte berechnen. Sie sind (da Korpus B das Zehnfache von Korpus A umfasst und bezüglich Teilkorpora gleich strukturiert ist) einfach zehnmal so hoch wie in Korpus A und entsprechen deshalb genau den beobachteten Werten. Korrekterweise folgt daraus $\chi^2 = 0$ und $p = 1$ ($df = 4$); die Verteilung unterscheidet sich also überhaupt nicht von jener in Korpus A.

Falls aber ein signifikanter Unterschied zu den erwarteten Werten feststellbar wäre, könnte wieder über einen paarweisen Vergleich festgestellt werden, welche Teilkorpora von den erwarteten Werten signifikant abweichen.

Würden wir in Korpus B aber die Werte Publikumspresse = 1 000, Bücher = 1 050, Internet/Wikipedia = 950, Gesprochenes = 7 000 und Sonstige Printmedien = 1 070 beobachten, würde sich auf der Basis der erwarteten Werte ein leicht signifikanter Unterschied von $p = 0{,}04$ in der Verteilung der Treffer auf die Teilkorpora im Vergleich zu Korpus A ergeben.

Bei einer ganz anderen Zusammensetzung des Untersuchungskorpus im Vergleich zu Korpus A wird die Kontingenztabelle der erwarteten Werte entsprechend angepasst: Tabelle 8 zeigt angenommene beobachtete Werte für ein Phänomen X in einem Korpus C.

| Korpus C beobachtet | X | ¬X | Total |
|---|---|---|---|
| Publikumspresse | 45 000 | 1 999 955 000 | 2 000 000 000 |
| Bücher | 11 000 | 499 989 000 | 500 000 000 |
| Internet/Wikipedia | 4 500 | 199 995 500 | 200 000 000 |
| Gesprochenes | 1 600 | 9 998 400 | 10 000 000 |
| Sonstige Printmedien | 9 000 | 399 991 000 | 400 000 000 |
| **Total** | **71 100** | **3 109 928 900** | **3 110 000 000** |

Tab. 8: Beobachtete Werte von Phänomen X in den Teilkorpora von Korpus C

Die erwarteten Werte für Tabelle 8 widerspiegeln nun die Verteilung der beobachten Werte von Tabelle 1 auf die Teilkorpora (Publikumspresse = 100, Bücher = 100, Internet/Wikipedia = 100, Gesprochenes = 700, Sonstige Printmedien = 100, Total = 1100), allerdings angepasst an die Trefferzahl von 71100 und die anderen Verhältnisse der Größen der Teilkorpora zueinander (Tabelle 9).[5]

| Korpus C erwartet | X | ¬X | Total |
|---|---|---|---|
| Publikumspresse | 44858 | 1999955142 | 2000000000 |
| Bücher | 11215 | 499988785 | 500000000 |
| Internet/Wikipedia | 4486 | 199995514 | 200000000 |
| Gesprochenes | 1570 | 9998430 | 10000000 |
| Sonstige Printmedien | 8972 | 399991.028 | 400000000 |
| **Total** | **71100** | **3109928900** | **3110000000** |

Tab. 9: Erwartete Werte von Phänomen X in den Teilkorpora von Korpus C auf der Basis der beobachteten Werte von X in Korpus A (vgl. Tabelle 1)

Der Chi-Quadrat-Test ergibt nun p = 0,26 ($\chi^2$ = 5,259, df = 4), also keine signifikante Abweichung von der Verteilung in Korpus A. Die Verteilung der Frequenzen in Tabelle C auf die Teilkorpora weichen kaum von der Verteilung in Korpus A ab.

## 3.3 Verteilung über das Gesamtkorpus

Die oben dargestellten statistischen Tests können verwendet werden, um eine unregelmäßige Verteilung über die Teilkorpora zu entdecken. Trotzdem kann mit diesen Tests noch nicht ausgeschlossen werden, dass innerhalb eines Teilkorpus eine sehr ungleichmäßige Verteilung der Treffer herrscht. Im Extremfall kann sich ein Phänomen auf nur einen Text (z.B. einen längeren Roman) beschränken und für die hohe Frequenz im entsprechenden Teilkorpus verantwortlich sein (vgl. dazu das Beispiel zu *frug* in Kapitel 1.2.2).

[5] Der erwartete Wert einer Zelle berechnet sich in zwei Schritten über eine Hilfstabelle A', die die Verteilung von Tabelle A in den Teilkorpusgrößen von Tabelle C widerspiegelt: $A'_E$ = A · Zeilensumme C : Zeilensumme A. Die eigentliche Kontingenztabelle für Korpus C basiert dann auf Tabelle A' und interpoliert die Werte daraus auf die Summe der Fälle für X in Tabelle C: $C_E$ = $A'_E$ · Spaltensumme C : Spaltensumme $A'_E$.

Auf dieses Problem der „Clumpiness" verweisen verschiedene Autoren (Church/Gale 1995; Kilgarriff 2001:107; Evert 2006; Gries 2008a). Es gibt verschiedene statistische Maße, um die Gleichmäßigkeit der Verteilung eines Phänomens auf ein Korpus zu messen, wobei dafür das Korpus immer in kleinere Einheiten, z.B. Texte oder beliebige andere Abschnitte, unterteilt wird. Gries (2008a) diskutiert statistische Maße, die die Streuung der Frequenzen in den jeweiligen Korpusteilen bemisst (z.B. Juilland et al.s *D*, Rosengrens *S* etc.) bzw. erwartete mit beobachteten Frequenzen in den Korpusteilen miteinander in Beziehung setzen.

Dabei zeigt sich, dass keines dieser Maße befriedigend ist, da sie meistens davon ausgehen, dass die einzelnen Korpusteile gleich groß sind, keine normalisierten Werte ausgeben oder zu über- bzw. unterempfindlich gegenüber Häufungsschwankungen in Korpusdaten sind (vgl. für Details Gries 2008a). Deshalb schlägt Gries ein Maß vor, das speziell für Korpusdaten geeignet ist: Gries DP ('Deviation of Proportions'). Es wird folgendermaßen berechnet:

- Für jedes Teilkorpus wird die **erwartete** relative Frequenz des Phänomens berechnet.
- Für jedes Teilkorpus wird die **beobachtete** relative Frequenz des Phänomens berechnet.
- Die Differenzen zwischen erwarteten und beobachteten Prozentwerten werden pro Teilkorpus berechnet und die Differenzen aller Teilkorpora summiert und durch 2 dividiert.
- In einem letzten Schritt wird der berechnete Wert DP zu DPnorm normiert, um zu ermöglichen, dass der Wert die theoretischen Maxima und Minima 1 und 0 erreichen kann: DP/(1-min(*s*)), wobei min(*s*) die relative Größe des kleinsten Teilkorpus ausdrückt.[6]

DPnorm ergibt immer einen Wert zwischen 0 und 1. Je näher der Wert bei 0 liegt, desto gleichmäßiger ist das Phänomen verteilt.

Zur Illustration dienen nachfolgende Berechnungen von verschiedenen Verteilungsmaßen für Recherchen im DeReKo. Es wurde mit unterschiedlichen Korpora gearbeitet: Das erste Korpus ('10%-Korpus') umfasst eine Zufallsauswahl von 10% der Anzahl Wörter des gesamten DeReKo und damit 374 521 682 Wörter und 1 547 513 Texte. In Ergänzung dazu wurde für einzelne Lexeme

[6] Die Formel für DPnorm ist gegenüber Gries (2008a) in Lijffijt/Gries (2012) um einen Fehler korrigiert worden.

auch im kompletten DeReKo recherchiert. Zudem wurden Teilkorpora von Pressetexten in Deutschland (D), Österreich (A) und der Schweiz (CH) erstellt, die jeweils alle entsprechenden Zeitungen aus dem DeReKo verwenden.[7]

Im DeReKo sind mehrere Texte zu „Dokumenten" zusammengefasst (z.B. alle Zeitungsartikel aus einem Monat einer Zeitung). Pro Dokument wird die Frequenz des Phänomens berechnet. Die Ergebnisse für verschiedene Maße werden mit einem R-Script[8] von Gries berechnet (Gries 2008a, 2009).[9] In den Tabellen 10 und 11 sind die Ergebnisse für Gries DP im Vergleich zu anderen Maßen angegeben.

| Suche | SD | VC | Chi-Quadrat | Juillands D | DPnorm | Frequenz/Mio. |
|---|---|---|---|---|---|---|
| *aber (10%)* | 203,84 | 0,79 | 26735,77 | 0,99 | 0,07 | 2105,10 |
| *aber* | 2068,84 | 0,88 | 245816,4 | 0,99 | 0,07 | 2110,07 |
| *diskutieren (10%)* | 11,30 | 1,05 | 6352526 | 0,98 | 0,16 | 88,35 |
| *dieses Jahres (10%)* | 5,07 | 1,14 | 7875384 | 0,98 | 0,27 | 36,55 |
| *diesen Jahres (10%)* | 0,98 | 2,19 | 3774208 | 0,95 | 0,59 | 3,68 |
| *parken (10%)* | 4,10 | 1,80 | 7711566 | 0,95 | 0,38 | 18,61 |
| *parkieren (10%)* | 0,80 | 5,02 | 8200104 | 0,91 | 0,92 | 1,30 |
| *parken CH* | 2,31 | 0,87 | 429,1861 | 0,94 | 0,30 | 1,36 |
| *parkieren CH* | 14,71 | 0,63 | 856,6168 | 0,97 | 0,16 | 14,62 |
| *fragte (10%)* | 4,63 | 1,37 | 21923,85 | 0,96 | 0,29 | 27,60 |
| *frug (10%)* | 0,15 | 10,70 | 5328926 | 0,77 | 0,98 | 0,11 |

**Tab. 10: Verschiedene Maße[10] zur Berechnung der Verteilung von Verben und anderen Phänomenen in unterschiedlichen Korpora[11]**

[7] Die Korpora umfassen die folgenden Zeitungen: Schweiz: St. Galler Tagblatt, Zürcher Tages-Anzeiger, Südostschweiz (1206430 Texte, 347497072 Wörter); Deutschland: alle im DeReKo verfügbaren deutschen Zeitungen ab 1980 (11261444 Texte, 2811488592 Wörter); Österreich: Burgenländische Volkszeitung, Die Presse, Kleine Zeitung, Neue Kronen-Zeitung, Niederösterreichische Nachrichten, Oberösterreichische Nachrichten, Salzburger Nachrichten, Tiroler Tageszeitung, Voralberger Nachrichten (3366977 Texte, 618878022 Wörter).

[8] Zu R vgl. Ihaka/Gentleman (1996) sowie Kapitel 3.4 („Implementierung in R").

[9] Vgl. http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/links.html (Stand: 2.8.2013).

[10] Standardabweichung (SD), Variationskoeffizient (VC), Chi-Quadrat-Test, Juilland et al.s D für unterschiedlich große Korpusteile, Gries DPnorm (vgl. Gries 2008a: 407, 415ff.).

[11] Der Zusatz '10%' bedeutet: 10%-Zufallsauswahl des DeReKo; Länderkürzel CH, D, A: Schweizer, deutsches und österreichische Pressekorpora aus dem DeReKo; ohne zusätzliche Angabe: komplettes DeReKo.

An DPnorm ist ersichtlich, dass *aber* mit einem Wert von 0,07 sehr gleichmäßig über die Dokumente verteilt ist. Dies gilt sowohl für die Recherche im Gesamtkorpus des DeReKo als auch in der 10%-Zufallsauswahl.[12] Etwas weniger gleichmäßig verteilt ist *diskutieren*.

Die beiden Wörter dienen uns im Vergleich zu den anderen Ausdrücken als Referenzpunkt, von denen – insbesondere bei *aber* – nicht erwartet wird, dass sie bezüglich irgendeines Faktors unregelmäßig im Korpus anzutreffen wären. Bei *diskutieren* mag es im Vergleich zu *aber* in bestimmten Texten aufgrund inhaltlicher Verwendungsbeschränkungen bereits zu etwas stärker gehäufter Verwendung kommen.

Wie verhalten sich die grammatisch interessanteren Ausdrücke zu diesen beiden Referenzpunkten? Alle sind weniger gleichmäßig verteilt als *aber* und *diskutieren*. Der Ausdruck *diesen Jahres* ist weniger gleichmäßig verteilt (DPnorm = 0,59) als *dieses Jahres* (DPnorm = 0,27).[13] Sehr ungleichmäßig ist der Helvetismus *parkieren* verteilt (DPnorm = 0,92), da das Lexem wahrscheinlich nur in Schweizer Quellen benutzt wird. Wenn die Suche auf die im Korpus verfügbaren Schweizer Zeitungen beschränkt wird, ergibt sich erwartungsgemäß ein sehr kleiner DPnorm-Wert von 0,16 (vgl. *parkieren CH* in Tabelle 10); in der Schweiz ist das Verb also ziemlich gleichmäßig in den Daten verteilt. Am ungleichmäßigsten ist der Ausdruck *frug* verteilt, eine heutzutage eindeutig standardferne Präteritalform (DPnorm = 0,98).

Die Ausdrücke *fragte* und *parken* sind mit einem DPnorm-Wert von 0,28 bzw. 0,38 relativ gleichmäßig verteilt, jedoch noch immer ungleichmäßiger als *aber* und *diskutieren*. Die Rangfolge der Suchausdrücke von sehr gleichmäßig bis nicht gleichmäßig (*aber* > *diskutieren* > *dieses Jahres* > *fragte* > *parken* > *diesen Jahres* > *parkieren* > *frug*) scheint plausibel zu sein; von den oben berechneten Maßen spiegelt nur der Variationskoeffizient diese Reihenfolge wider. Die Standardabweichung und der Chi-Quadrat-Wert sind stark von der absoluten Frequenz der Treffer abhängig, deshalb weist *aber* den höchsten Wert auf. Juilland et al.s D zeigt sich wenig empfindlich gegenüber den unterschiedlichen Frequenzschwankungen.

[12] Allerdings weichen die Standardabweichung und der Chi-Quadrat-Wert bei *aber* in den beiden Korpora massiv voneinander ab, was zeigt, dass diese beiden Maße definitionsgemäß stark von der Korpusgröße abhängig sind.

[13] Obwohl beide Varianten korrekt sind, urteilt die „Grammatik in Fragen und Antworten" dazu: „Aus sprachtheoretischer Sicht lässt sich die eingangs gestellte Frage nicht eindeutig beantworten. Wer sich damit nicht zufriedengeben will, mag eine praktische Entscheidung treffen: Kritik zieht man sich allenfalls zu, wenn man *diesen* verwendet. Die Form *dieses* wird in Verbindung mit jedem Nomen im Genitiv überall und jederzeit fraglos akzeptiert." (vgl. z.B. Strecker 2010a).

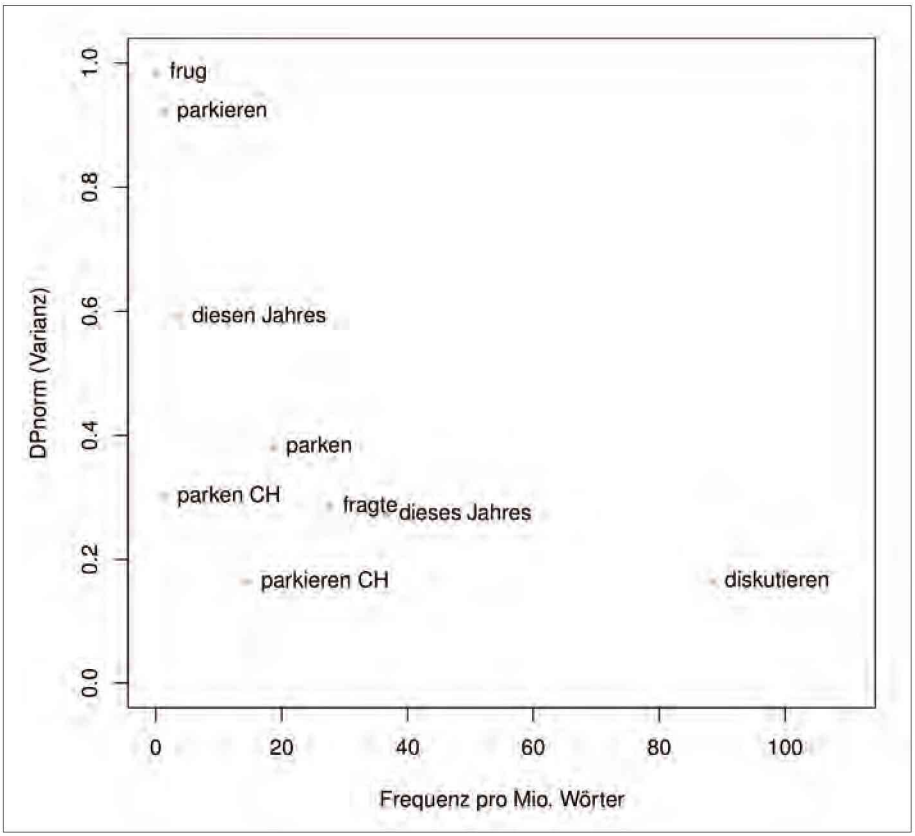



**Abb. 1: Plot einiger Lexeme aus Tabelle 10 auf den Achsen DPnorm und relativer Frequenz im Korpus**

Abbildung 1 bildet die Ausdrücke aus Tabelle 10 als Korrelation zwischen DPnorm und Frequenz im Korpus ab.[14] Die Visualisierung macht deutlich, dass Wörter, die im unteren Bereich der Grafik liegen, eher zum Standard gehören, während die Wörter im oberen Bereich eher standardfern sind. Zudem sind Wörter im linken Bereich niedrig- und im rechten Bereich hochfrequent im Korpus. Allerdings ist eine direkte Abbildung des DP-Wertes auf Standardnähe oder -ferne problematisch: Gemäß Abbildung 1 wäre *fragte* etwas standardferner als *diskutieren*. Man würde aber wohl beide Wörter gleichermaßen zum Standard zählen wollen. Die gegenüber *fragte* zu testende Variante wäre aber auch nicht *diskutieren*, sondern z.B. *frug*, das über ein Tertium Compa-

[14] Auf die Darstellung von *aber* in der Grafik wurde aus Gründen der Übersichtlichkeit verzichtet, da die Frequenz dieses Ausdrucks sehr viel höher ist als die der anderen Ausdrücke.

rationis (vgl. Kapitel 1.2.1) mit *fragte* verglichen werden kann. Die DP-Werte von *fragte* und *frug* oder eben *parkieren* und *parken* etc. unterscheiden sich stark, so dass klar ist, welche Variante eher als standardfern gewertet werden kann.

| Suche | SD | VC | Chi-Quadrat | Juillands D | DPnorm | Frequenz/ Mio. |
|---|---|---|---|---|---|---|
| *Antragsteller (10%)* | 0,93 | 2,14 | 5 557 407 | 0,93 | 0,61 | 3,55 |
| *Antragssteller (10%)* | 0,16 | 8,22 | 4 083 908 | 0,82 | 0,97 | 0,16 |
| *Antragsteller CH* | 2,54 | 0,98 | 430 | 0,94 | 0,33 | 1,63 |
| *Antragssteller CH* | 0,77 | 1,78 | 231 | 0,88 | 0,65 | 0,27 |
| *Schweinebraten (10%)* | 0,29 | 4,72 | 4 732 783 | 0,89 | 0,92 | 0,50 |
| *Schweinsbraten (10%)* | 0,25 | 5,65 | 5 572 | 0,86 | 0,95 | 0,36 |
| *Schweinebraten* | 1,01 | 1,94 | 6 610 378 | 0,95 | 0,59 | 0,47 |
| *Schweinsbraten* | 1,21 | 3,16 | 10 725 | 0,94 | 0,77 | 0,34 |
| *Schweinebraten D* | 1,20 | 1,49 | 4 404 427 | 0,96 | 0,53 | 0,57 |
| *Schweinsbraten D* | 0,52 | 2,99 | 4 302 002 | 0,91 | 0,84 | 0,12 |
| *Schweinebraten A* | 0,57 | 2,81 | 950 | 0,87 | 0,80 | 0,19 |
| *Schweinsbraten A* | 1,35 | 1,60 | 874 957 | 0,93 | 0,49 | 1,53 |
| *Schweinebraten CH* | 0,59 | 2,20 | 301 546 | 0,83 | 0,77 | 0,17 |
| *Schweinsbraten CH* | 0,54 | 2,67 | 352,52 | 0,80 | 0,82 | 0,12 |

**Tab. 11: Verschiedene Maße[15] zur Berechnung der Verteilung von Komposita in unterschiedlichen Korpora[16]**

In Tabelle 11 sind die Werte für zwei Komposita aufgeführt: *Antragsteller* und *Schweinebraten* jeweils mit den *s*-Fugen-Varianten *Antragssteller* und *Schweinsbraten*.

Komposita erscheinen im Korpus insgesamt deutlich seltener und sind dadurch auch von vornherein nicht so gleichmäßig verteilt. Das unumstritten standarddeutsche *Antragsteller* ist sogar etwas ungleichmäßiger verteilt (DPnorm = 0,61) als das tendenziell nur standardnahe *diesen Jahres* aus Tabelle 10. Dennoch ist der Abstand zum eher als standardfern einzustufenden *Antragssteller* (DPnorm = 0,97) noch sehr deutlich (lediglich in der Schweiz ist dieses etwas gleichmäßiger verteilt – DPnorm = 0,65). Aufschlussreich erscheint also

[15] Standardabweichung (SD), Variationskoeffizient (VC), Chi-Quadrat-Test, Juilland et al.s D für unterschiedlich große Korpusteile, Gries DPnorm (vgl. Gries 2008a: 407/415ff.).

[16] Der Zusatz '10%' bedeutet: 10%-Zufallsauswahl des DeReKo; Länderkürzel CH, D, A: Schweizer, deutsches und österreichische Pressekorpora aus dem DeReKo; ohne zusätzliche Angabe: komplettes DeReKo.

auch hier, wie bereits oben bei Abbildung 1 dargestellt, vor allem der Vergleich von Komposita und deren Varianten untereinander, was DPnorm für die Ausdrücke *Schweinebraten* und *Schweinsbraten* besonders gut illustrieren kann: *Schweinebraten* ist im Gesamtkorpus ähnlich verteilt (DPnorm = 0,59) wie *Antragsteller* und gleichmäßiger als *Schweinsbraten* (DPnorm = 0,77). Der Abstand zwischen den beiden Varianten ist in den Quellen aus Deutschland noch deutlicher. In den österreichischen Quellen kehrt sich das Verhältnis zwischen den Varianten aber um – der Ausdruck *Schweinsbraten* erreicht hier sogar den niedrigsten DPnorm-Wert in Tabelle 11 überhaupt. In den Schweizer Quellen schließlich liegen die DPnorm-Werte für beide Varianten nah beieinander, wobei *Schweinebraten* etwas gleichmäßiger verteilt erscheint.

Abbildung 2 bildet die Ausdrücke aus Tabelle 11 als Korrelation zwischen DPnorm und Frequenz im Korpus ab. Die Rangfolge der Ausdrücke von gleichmäßig (unten) bis nicht gleichmäßig (oben) *Schweinsbraten* A > *Schweinebraten* D > *Schweinebraten* > *Antragsteller* > *Schweinsbraten* > *Schweinebraten* CH > *Schweinebraten* A > *Schweinsbraten* CH > *Schweinsbraten D* > *Antragssteller* scheint erneut plausibel in Bezug auf die Überlegungen zur Standardzugehörigkeit zu sein, da der Ausdruck *Schweinsbraten* bekanntlich insbesondere im süddeutschen Raum Verwendung findet.[17]

Die in Abbildung 2 dargestellten Werte für *Schweinebraten* und *Schweinsbraten* beziehen sich nicht auf das 10%-Korpus, sondern auf das komplette DeReKo. Anders als bei *aber* in Tabelle 10 hat sich nämlich bei niedrigfrequenten Phänomenen gezeigt, dass mitunter der DPnorm-Wert je nach Korpusgröße stark schwankt. *Schweinebraten* ergibt im 10%-Korpus einen DPnorm-Wert von 0,92 (statt 0,59 im gesamten DeReKo) und *Schweinsbraten* von 0,95 (statt 0,77). Zwar ist es auch im 10%-Korpus so, dass *Schweinebraten* etwas gleichmäßiger verteilt ist als *Schweinsbraten*, der Unterschied ist aber minimal und die Werte im Vergleich zu den anderen Phänomenen sehr hoch. Warum dies so ist, muss noch genauer untersucht werden. Jedoch scheint es gerade bei niedrigfrequenten Phänomenen wichtig zu sein, mit möglichst großen Datenmengen zu arbeiten.

[17] Lediglich die etwas gleichmäßigere Verteilung von *Schweinebraten* in den Schweizer Quellen kann linguistisch erklärungsbedürftig erscheinen.

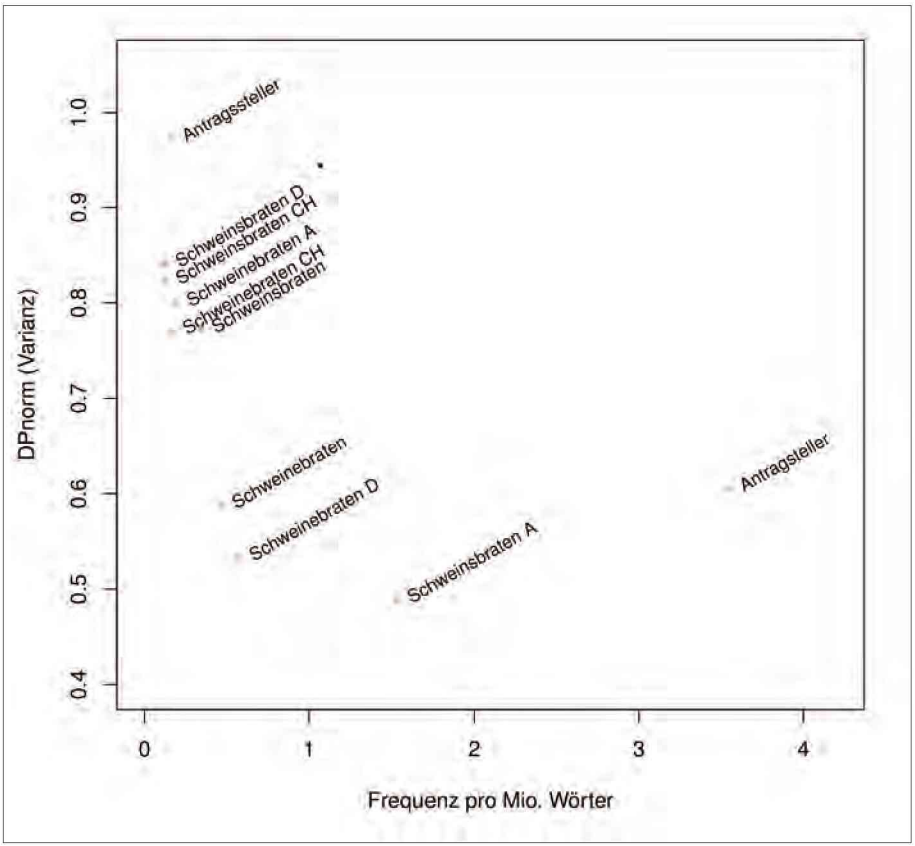



Abb. 2: Plot einer Auswahl von Lexemen aus Tabelle 11 auf den Achsen DPnorm und relativer Frequenz im Korpus

Es kann an dieser Stelle keine ausführliche Evaluation der unterschiedlichen Maße erfolgen. Die Studien von Gries zu unterschiedlichen Lexemen im British National Corpus (Gries 2008a : 419ff.) sind jedoch ermutigend, Gries DP als Kontrollmaß anzuwenden, um zu überprüfen, ob ein bestimmtes Phänomen in bestimmten Dokumenten gehäuft vorkommt. Die weiteren Studien im Projekt werden den Nutzen des Maßes überprüfen.

## 3.4 Implementierung in R

Mit der Statistikumgebung R (Ihaka/Gentleman 1996) steht ein Werkzeug zur Verfügung, um die vorgestellten statistischen Berechnungen automatisiert durchzuführen. Die Berechnungen werden zu Scripts zusammengefasst, die in

die KoGra-DB (vgl. Kapitel 2) eingebettet werden können und über das Web-Interface, über das die Korpusdatenbank bedient wird, aufgerufen werden.

### 3.4.1 Chi-Quadrat-Tests für die Frequenzvergleiche zwischen Teilkorpora

Um den Chi-Quadrat-Test durchzuführen, kann in R die Funktion `chisq.test()` benutzt werden. Sie liest die Daten der beobachteten Werte als Matrix ein und gibt den Chi-Quadrat-Wert und die Wahrscheinlichkeit für die Nullhypothese p aus. Mit folgenden Befehlen wird die Matrix erstellt und die Statistik berechnet (ausgehend von den Daten aus Tabelle 1 ohne Totale):

```
1  x <- matrix(
   c(100,100,100,700,100,3999900,3999900,3999900,3999300,
   3999900),5,2
   );
2  dimnames(x) <- list(
   c("PubPr","Bücher","Int/Wiki","Gespr","Sonst"),
   c("X","-X")
   );
3  chisq.test(x);
4  Pearson's Chi-squared test
   data:  x
   X-squared = 1309.163, df = 4, p-value < 2.2e-16
```

In Zeile 1 wird eine Matrix mit den Daten der Kontingenztabelle erstellt, die 5 Zeilen und 2 Spalten aufweist.[18] In Zeile 2 werden den Zeilen und Spalten der besseren Übersichtlichkeit wegen sprechende Namen hinzugefügt. In Zeile 3 erfolgt schließlich der Chi-Quadrat-Test auf der Basis der Daten. Zeile 4 zeigt die Ausgabe des Tests.

Im Anschluss werden noch die Assoziationsstärke Cramérs V berechnet (Zeile 5, Resultat Zeile 6) und bei Tabellen, in denen ein signifikanter Frequenzunterschied feststellbar ist, ein paarweiser Vergleich gemacht (Zeile 7, Resultat Zeile 8):

```
5  sqrt(
         chisq.test(x)$statistic
         /
         sum(x)*(min(dim(x))-1)
   );
```

[18] Natürlich können die Werte auch als Tabelle mit read.table() eingelesen werden.

```
6   X-squared
    0.008090621

7   pairwise.prop.test(x, p.adjust.method="bonferroni");

8   Pairwise comparisons using Pairwise comparison of
    proportions
    data: x
              PubPr  Bucher Int/Wiki Gespr
    Bucher    1      -      -        -
    Int/Wiki  1      1      -        -
    Gespr     <2e-16 <2e-16 <2e-16   -
    Sonst     1      1      1        <2e-16
    =     P value adjustment method: bonferroni
```

Für Fragestellung 3 (vgl. Kapitel 3.2.3) muss die R-Funktion für die Berechnung des Chi-Quadrat-Tests zusätzlich mit den Angaben der angepassten erwarteten Werte aufgerufen werden (ausgehend von den Daten aus Tabellen 8 und 9):

```
9   x <- c(45000, 11000, 4500, 1600, 9000);

10  p <- c(44858, 11215, 4486, 1570, 8972);

11  chisq.test(x, p = p, rescale.p = TRUE);
```

Die benötigten R-Befehle werden von einem PHP-Script erzeugt, sobald der Benutzer in der *KoGra-DB* die statistische Standardauswertung aufruft. Die Resultate der Berechnungen werden als HTML-Seite an den Benutzer zurückgegeben (vgl. Abbildung 3). Neben den Zahlenwerten werden zudem die relativen Frequenzen als Balkendiagramme visualisiert.

### 3.4.2 DPnorm zum Messen der Verteilung über das Gesamtkorpus

DPnorm und weitere Maße, die die Verteilung eines Phänomens über ein Gesamtkorpus hinweg audrücken, lassen sich ebenfalls in R mit einem Script berechnen. Ausgangsbasis ist eine Tabelle, die für jedes Teilkorpus die Anzahl enthaltener Wörter und Treffer aufführt. Da nicht für jedes Phänomen die Einheit Wort sinnvoll ist (z.B. bei syntaktischen Strukturen), werden gegebenenfalls auch die Angaben für Sätze und Texte angegeben. Die Daten haben damit die Struktur wie in Tabelle 12 angedeutet.

Tabelle MEDIUM

**Daten:**

```
                  [,1]      [,2]
Publikumspresse 14704 244245810
Buecher             5   7133328
Internet           21   4183210
Gesprochenes        5    193126
```

**Chi-Quadrat-Test:**

```
        Pearson's Chi-squared test

data:  data
X-squared = 633.6426, df = 3, p-value < 2.2e-16
```

**Phi Assoziationsstärke:**

```
   X-squared
0.001573973
```

**Paarweiser Vergleich:**

```
        Pairwise comparisons using Pairwise comparison of proportions

data:  data

            Publikumspresse Buecher Internet
Buecher     < 2e-16         -       -
Internet    < 2e-16         5.8e-05 -
Gesprochenes 0.4347         3.7e-16 0.0082

P value adjustment method: bonferroni
```

**Relative Frequenzen (pro Mio. Sätze):**

```
Publikumspresse        Buecher       Internet    Gesprochenes
     60.2016469      0.7009351      5.0200683      25.8898336
```

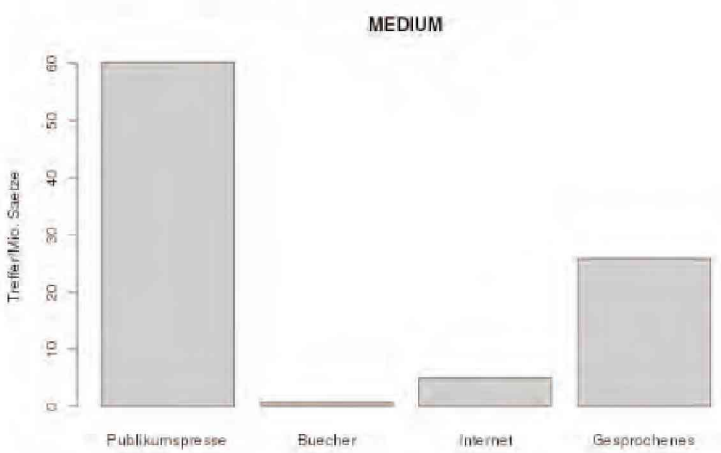



Abb. 3: Browser-Ausgabe der statistischen Tests zu Frequenzvergleichen

Tabelle DOMAENE

**Daten:**

```
                  [,1]      [,2]
Fiktion              0    354223
Kultur            4031 103480892
Mensch             326   6165203
Politik           8761 105359486
Technik           1219  17412207
unklassifizierbar  398  22983463
```

**Chi-Quadrat-Test:**

```
        Pearson's Chi-squared test

data:  data
X-squared = 2535.101, df = 5, p-value < 2.2e-16
```

**Phi Assoziationsstärke:**

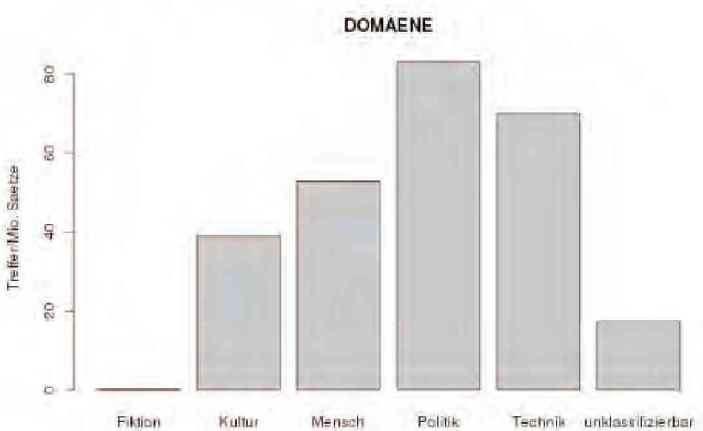
```
   X-squared
0.000148275
```

**Paarweiser Vergleich:**

```
        Pairwise comparisons using Pairwise comparison of proportions

data:  data

                  Fiktion Kultur  Mensch  Politik Technik
Kultur            0.00516 -       -       -       -
Mensch            0.00039 1.8e-06 -       -       -
Politik           1.4e-06 < 2e-16 1.0e-14 -       -
Technik           1.6e-05 < 2e-16 0.00011 2.8e-07 -
unklassifizierbar 0.34657 < 2e-16 < 2e-16 < 2e-16 < 2e-16

P value adjustment method: bonferroni
```

**Relative Frequenzen (pro Mio. Sätze):**

```
     Fiktion          Kultur          Mensch          Politik
     0.00000        38.95405        52.87742        83.15340
     Technik unklassifizierbar
    70.00836        17.31680
```



| Dokument | Texte Total | Wörter Total | Treffer (Texte) | Treffer |
|---|---|---|---|---|
| A00/APR | 606 | 180 667 | 0 | 0 |
| A00/AUG | 709 | 205 668 | 1 | 2 |
| A00/DEZ | 233 | 72 849 | 0 | 0 |
| A00/FEB | 761 | 227 292 | 0 | 0 |
| A00/JAN | 785 | 223 905 | 0 | 0 |
| A00/JUL | 624 | 188 156 | 0 | 0 |
| A00/JUN | 671 | 207 414 | 0 | 0 |
| A00/MAI | 886 | 261 496 | 0 | 0 |
| ... | | | | |

Tab. 12: Treffer für *diesen Jahres* in Dokumenten; Tabelle als Basis für die Berechnung von DPnorm

Nun wird diese Tabelle in R eingelesen und DPnorm neben vielen weiteren Maßen zur Verteilung berechnet:

```
1 data <-read.table(<Dateipfad>, header=TRUE, sep="\t",
  quote="", comment.char="") ;
2 data_prob_texte <- data$Texte_Total/sum
  (as.numeric(data$Texte_Total)) ;
3 data_prob_tokens <- data$Tokens_Total/sum
  (as.numeric(data$Tokens_Total)) ;
4 source("http://www.linguistics.ucsb.edu/faculty/
  stgries/research/dispersion/_dispersions2.r") ;
5 griesDPTokens <- dispersions2(data[[col_treffer]],
  data_prob_tokens) ;
6 griesDPTexte <- dispersions2(data[[col_texte]],
  data_prob_texte) ;
7 griesDPTokens ;
8 griesDPTexte ;
```

In Zeile 1 werden die Daten eingelesen und in den Zeilen 2 und 3 die bezüglich der Gesamtzahl von Wörtern und Texten relativen Teilkorpusgrößen berechnet. In Zeile 4 wird das Script zur Berechnung der statistischen Verteilungsmaße geladen und in Zeilen 5 und 6 die Werte gesondert für die Wort- und die Textfrequenzen berechnet und in Zeilen 7 und 8 ausgegeben.

Für das Recherchebeispiel *aber* (vgl. Tabelle 10) wird folgende Ausgabe (Basis Wörter) erzeugt (Abbildung nur auszugsweise unter Auslassung nicht erwähnter Maße):

```
$'observed overall frequency'
[1] 788405

$'sizes of corpus parts / corpus expected proportion'
 [1] 4.823940e-04 5.491484e-04 1.945121e-04 6.068861e-04
     5.978426e-04
 [6] 5.023901e-04 5.538104e-04 6.982132e-04 6.603303e-04
     6.376827e-04
[11] 5.687227e-04 7.117986e-04 1.432227e-04 3.854196e-04
     3.185503e-04
[16] 3.224433e-04 4.942411e-04 3.809980e-04 2.054354e-04
     6.915941e-04
...

$'relative entropy of all sizes of the corpus parts'
[1] 0.9546112

$range
[1] 3045

$maxmin
[1] 1084

$'standard deviation'
[1] 203.8447

$'variation coefficient'
[1] 0.7932414

$'Chi-square'
[1] 26735.77

$'Juilland et al.'s D (not requiring equally-sized corpus parts)'
[1] 0.9941908

$'Deviation of proportions DP (normalized)'
[1] 0.0702684
```

Im regulären Rechercheprozess werden nicht alle Maße ausgegeben, sondern die Ausgabe auf DPnorm eingeschränkt.

## 3.5 Ausblick

Ein „statistischer Schnelltest" birgt die Gefahr, uninformiert nicht die adäquaten Testverfahren auf die Daten anzuwenden. Zudem können nur eine Reihe von Standardbedingungen berücksichtigt werden: Die skizzierten Verfahren sind dafür gemacht, die Verteilung eines Phänomens in den definierten Teilkorpora zu untersuchen und ein Phänomen mit einem oder mehreren weiteren Phänomenen zu vergleichen. Sollen andere Teilkorpora oder mehrere Phänomene miteinander verglichen werden, müssen über diesen Standardtest hinausgehende erweiterte statistische Verfahren angewandt und/oder die Korpusdaten speziell dafür aufbereitet werden.

Es ist deshalb klar, dass der Rechercheprozess fortwährend statistisch begleitet werden muss. Das ermöglicht es auch, die statistischen Verfahren anzupassen, wenn sich erweisen sollte, dass sie den Bedürfnissen nicht genügen.

Ein weiterer Ausbau ist insbesondere bezüglich der Integration in die *KoGra-DB* geplant: Bisher können nur die Frequenzvergleiche direkt aus der *KoGra-DB* heraus aufgerufen werden. Die Berechnungen zur Gleichmäßigkeit der Verteilung (DPnorm) bedingen als Eingabedaten nicht nur Kontingenztabellen, sondern darüber hinaus auch Trefferlisten, aufgegliedert nach kleineren Einheiten (Dokumente, Texte etc.). Die Verarbeitung solcher großen Datenmengen ist jedoch nicht trivial (vgl. dazu auch Kapitel 2).

Sobald eine derart erweiterte Parameterübergabe realisiert ist, können die beiden Testgruppen besser verzahnt werden: So wird es möglich, dass die Berechnung von DPnorm automatisch separat pro Teilkorpus erfolgt und in Kombination mit den Signifikanztests zu den Frequenzunterschieden noch aussagekräftigere Analysen möglich sind.

# 4. Verlässlichkeit und Brauchbarkeit grammatischer Annotation

Untersuchungen zu Varianz und Variation grammatischer Strukturen, die auf sehr großen Korpora basieren, sind ohne maschinelle Unterstützung von Taggern, die automatisiert grammatische Klassifikationen des gesamten Wortmaterials vornehmen, kaum zu denken. Laut Studien zum POS-Tagging für deutsche Texte genügen die besten Tagger bereits hohen Ansprüchen. Von 93- bis 98-prozentiger Verlässlichkeit ist die Rede. Zu bedenken ist aber, dass das Gros deutscher Wörter einigen wenigen Klassen zugeordnet werden kann, die vergleichsweise eindeutig zu fassen sind. Erweisen sich die Tagger in diesem Bereich als leistungsfähig, kann also die sehr gute Durchschnittsverlässlichkeit manche Schwäche bei Wortklassen mit wenigen Elementen verdecken. Die Verlässlichkeit der Tagger soll daher exemplarisch im Hinblick auf Wortformen *meine/Meine*, die Präpositionen im Allgemeinen und die Wortformen *bar, eingangs fernab, halber, jenseits, laut* überprüft werden. Auf dem Prüfstein stehen TreeTagger und Connexor, mit denen die Korpora des Instituts für Deutsche Sprache annotiert sind. Kritisch betrachtet werden diese Tagger u.a. im Hinblick auf falsche Zuordnungen für ein gesuchtes Phänomen (*false positives*) und auf nicht vorgenommene Zuordnungen (*false negatives*). Die Untersuchung mündet in ernüchternden Erkenntnissen, aber auch in ersten Ideen zur sinnvollen Nutzung getaggter Korpora trotz ihrer echten und vermeintlichen Fehler.

## 4.1 Anlass und Rahmen der Untersuchung

Sucht man Aussagen über grammatische Strukturen und grammatische Varianz bzw. Variation im heutigen Deutsch gestützt auf sehr große Textkorpora zu evaluieren, erkennt man schnell, dass dies ohne maschinelle Unterstützung kaum zu realisieren ist. Man muss einen Weg finden, zumindest auf der elementaren Ebene der Wörter automatisiert Klassifikationen in der Art vorzunehmen, wie sie von Grammatikern schon seit den Zeiten von Dionysius Thrax[1] erarbeitet werden. Mit so genannten Taggern, die solche Klassifikationen auf der Basis entsprechender Algorithmen automatisch vornehmen können, scheinen ideale Lösungen für dieses Problem gefunden, die über eine bloße Evaluation bereits gegebener Aussagen zu grammatischen Strukturen

[1] Dionysios Thrax lebte vermutlich im 2. Jahrhundert v. Chr. und verfasste die erste griechische Grammatik, wobei er Überlegungen auswertete, die in den vorangegangenen Jahrhunderten zu Sprache und Grammatik angestellt worden waren.

hinaus sogar erlauben könnten, Strukturen zu entdecken, die Grammatikern bislang nicht in den Blick geraten sind.

Nun verwenden Grammatiker in aller Regel große Mühen darauf, die Wortklassen, von denen sie ausgehen, sorgfältig zu definieren, und sie lassen sich dabei auch nicht von Fehlern irritieren, wie sie sich bei der maschinellen Erfassung von Texten unterschiedlichster Ausgangsformate fast unvermeidlich ergeben. Vor allem aber legen sie – zumindest in Grammatiken mit wissenschaftlichem Anspruch – auch ihre Entscheidungsgründe offen, was kompetente Nutzer der Grammatik prinzipiell in die Lage versetzen kann, in beliebigen Texten allen Wörtern manuell die entsprechende Wortklasse zuzuordnen.

Bei maschinellem Tagging erfolgt die Zuweisung von Wortklassen schnell und selbstständig, doch der enorme Zeitgewinn hat einen hohen Preis: Die Kriterien, nach denen die Zuweisung erfolgt, bleiben – zumindest für jene, die nicht an der Konzeption des genutzten Taggers beteiligt waren – weitgehend undurchsichtig. Sie können sich nur durch die Hintertür einer Analyse der Ergebnisse ein Bild davon machen, welchen Regeln das Tagging folgt. Deshalb wird man als klassisch geschulter Grammatiker erst einmal zu überprüfen suchen, ob und in wie weit man sich auf deren Leistung verlassen kann, bevor man sich daran macht, mit einem oder mehreren Taggern zu arbeiten.

Schenkt man den Aussagen über die Qualität der auf dem Markt verfügbaren Taggern Glauben, scheinen zumindest die Besten unter ihnen bereits hohen Ansprüchen zu genügen. In den wenigen Untersuchungen zur Verlässlichkeit des POS-Taggings[2] für deutsche Texte ist von 93 bis 98% die Rede.[3] Doch bevor man sich von solch hohen Zahlen beeindrucken lässt, sollte man genauer betrachten, wie sie zustande gekommen sein könnten. Schon ein kurzer Blick auf den deutschen Wortschatz zeigt, dass das Gros deutscher Wörter einigen wenigen Wortklassen zugeordnet werden kann, die vergleichsweise eindeutig zu fassen sind. Erweist sich ein Tagger hinsichtlich dieser Klassen als leistungsfähig, wirkt sich das mittelbar auch auf seine Leistungsfähigkeit im Durchschnitt aller Wortklassen aus. Schwächen bei Wortklassen mit sehr viel weniger Elementen senken den Durchschnittswert dann nur wenig.

Ob und inwieweit Schwächen bei Wortklassen wie Artikel, Konnektoren oder Präpositionen die Brauchbarkeit eines Taggers wesentlich beeinträchtigen, hängt freilich in erster Linie davon ab, was man mit ihm erreichen will. Was sich im Hinblick auf eine Evaluation grammatischer Regeln als fatal erweisen

[2] POS: *Part of Speech.*

[3] Siehe hierzu Belica et al. (2009).

könnte, mag für ein effizientes Durchsuchen riesiger Textmengen unter rein sachlichen Aspekten zu vernachlässigen sein. Stellt man in Rechnung, dass die Nutzung von Taggern im Rahmen grammatischer Studien schon aus ökonomischen Gründen eher als ausgesprochen nachgeordnet gelten kann, sollte man sich zum einem hüten, allzu große Ansprüche an deren Leistung in Sachen grammatischer Analysen zu richten, zum anderen davon ausgehen, dass man die Tagger unter dem Aspekt einer verlässlichen Zuordnung von Wortklassen eigens zu überprüfen hat.

### Exkurs: Wie funktioniert ein POS-Tagger?

Um einen Wortarten-Tagger evaluieren und die Ergebnisse verstehen zu können, ist es notwendig zu wissen, wie Tagger arbeiten. In der Computerlinguistik entwickelten sich zwei grundsätzliche Verfahren des Taggens (vgl. Carstensen et al. 2001: 373ff.; Lemnitzer/Zinsmeister 2006: 60ff.):

**1. Regelbasiertes Tagging:** Der Computer verlässt sich beim Tagging auf ein umfangreiches Regelwerk, ein **Taggerlexikon**, das von Hand erstellt wurde und in dem festeht, welche Wörter welchen Wortarten angehören (gegebenenfalls in Abhängigkeit zum Kontext).

**2. Stochastisches Tagging:** Die Basis bei diesem Verfahren ist ebenfalls ein Taggerlexikon. Es wird jedoch zusätzlich ein sogenanntes Trainingskorpus erstellt, das von Hand getaggt wurde. Nun kann mit statistischen Methoden die Wahrscheinlichkeit von bestimmten Wortartenabfolgen berechnet werden. Dadurch entstehen auch Regeln, die jedoch auf Wahrscheinlichkeitswerten beruhen und beispielsweise besagen, dass in *der große Bär* das Token *der* wahrscheinlich kein Relativpronomen, sondern ein bestimmter Artikel ist, da dieser oft in der Kombination Artikel-Adjektiv-Nomen vorkommt. Die Qualität des Taggers ist deshalb stark vom verwendeten Trainingskorpus abhängig. Je größer dieses ist und je eher es den Daten ähnelt, die getaggt werden sollen, desto besser wird der Tagger arbeiten. Mit speziell angepassten Trainingskorpora kann die Taggingqualität verbessert werden.

Heute scheinen stochastische Tagger verbreiteter zu sein als regelbasierte. So arbeitet der sehr verbreitete TreeTagger (Schmid 1994) beispielsweise statistisch. Bei kommerziellen Produkten wie dem Connexor Machinese- und dem XEROX-Tagger ist die genaue Funktionsweise nicht öffentlich bekannt.

Stochastische Tagger haben den Vorteil, sehr robust zu sein: Sie können z.B. auch die Wortart von Wörtern erkennen, die nicht im Taggerlexikon enthalten sind, solang deren syntaktische Verwendung normgerecht ist und damit oft im Trainingskorpus angetroffen wurde. Doch gerade bei Untersuchungen zu Randbereichen des Standards ist dieses Verhalten problematisch und ein stochastischer Tagger tendiert dann zu mehr Fehlern. Allerdings gilt das auch für regelbasierte Tagger, wenn die entsprechenden Regeln nicht erfasst wurden.

## 4.2 Exemplarische Überprüfung

Will man als Grammatiker die Leistungsfähigkeit eines Taggers überprüfen, muss man sich zunächst darüber klar werden, nach welchen Kriterien diese erfolgen sollte. Durchaus verfehlt wäre etwa der Anspruch, der Tagger habe den in einem Korpus auftretenden Wortformen genau die Kategorien zuzuordnen, die man ihnen bei manueller Annotation auf der Basis der eigenen Grammatiktheorie zuordnete, denn Tagger verhalten sich in dieser Hinsicht nicht anders als die lieben Kollegen, die oft genug auch andere Wortklassen mit anderen Extensionen ansetzen als man selbst. Ob ein Tagger brauchbare und vor allem verlässliche Ergebnisse liefert, hängt nur bedingt davon ab, welche Basiskategorien er verwendet und welche Aufteilungen er dabei vornimmt. So führt etwa der TreeTagger[4] Ausdrücke wie *meine*, *deine*, *seine*, die man auf der Grundlage der „Grammatik der deutschen Sprache"[5] (GDS) als Possessivartikel bezeichnen würde, unter der Bezeichnung *PPOSAT* (*attribuierendes Possessivpronomen*). Sind derartige Unterschiede erst einmal erkannt, beeinträchtigen sie geplante Recherchen nicht weiter.

Selbst wenn ein Tagger eine Klasse von Wortformen, die man selbst derselben Wortklasse zurechnen würde, auf mehrere Klassen aufteilt oder umgekehrt mehrere solcher Klassen zu einer zusammenfasst, macht ihn dies nicht unbedingt unbrauchbar für Recherchen, die an sich von einer anderen Kategorisierung ausgehen. Solange der Tagger Zuordnungen vornimmt, die in die erwünschten Zuordnungen umgerechnet werden können, bleiben Recherchen möglich, die immer noch weit weniger mühsam sind, als solche, die sich nur auf reguläre Ausdrücke[6] stützen können, weil er zusätzliche Variablen bereitstellt, die kombiniert werden können, um gezielt nach Ausdruckssequenzen bestimmter formaler Strukturen zu suchen. Im Rahmen einer Evaluation grammatischer Regelhypothesen ist die Leistung eines Taggers deshalb allein danach zu bewerten, ob dieser – direkt, aber auch über „Workarounds" – erlaubt, eindeutig, effizient und möglichst vollständig die Phänomene aufzuspüren, deren Auftreten bzw. Ausbleiben man untersuchen will.

[4] Entwickelt am Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart.

[5] Zifonun et al. (1997), im Weiteren kurz GDS.

[6] Ein regulärer Ausdruck (engl. *regular expression*, Abk. *RegExp* oder *Regex*) ist eine Zeichenkette, mit der sich Mengen bzw. Untermengen von Zeichenketten über bestimmte syntaktische Regeln erfassen lassen.

Kritisch zu bewerten ist die Leistung eines Taggers erst, wenn er

- signifikant viele falsche Zuordnungen für ein gesuchtes Phänomen vornimmt (so genannte *false positives*).
- zahllose Zuordnungen nicht vornimmt, die er hätte vornehmen müssen, wenn die Annotation korrekt erfolgt wäre (sogenannte *false negatives*).

Ob ein Tagger in der einen oder anderen Weise fehlerhaft arbeitet, ist freilich immer nur relativ zu traditionellen Formen der Wortartbestimmung zu beurteilen, denn der Tagger arbeitet so, wie er eben arbeitet. Erkennt er etwa eine Wortform als Präposition, dann ist sie nach seinen Kriterien eine Präposition. Sind wir mit seinem Ergebnis nicht zufrieden, so allein deshalb, weil wir dieses abgleichen mit Ergebnissen einer von erfahrenen Grammatikern vorgenommen Analyse. Ein solcher Vergleich ist, sofern man berücksichtigt hat, was bereits als tolerable Abweichung von eigenen Klassifikationen genannt wurde, durchaus angebracht, denn man kann – etwa im Fall der Präpositionen – sehr wohl davon ausgehen, dass deren vollständige und korrekte Erkennung – ganz im Sinn einer klassisch grammatischen Definition – beabsichtigt war und deshalb Zuordnungen, die dieser Definition nicht genügen, nicht einfach als alternative Kategorisierungen hinzunehmen sind.

Gerade im Fall der Präpositionen sind die Kriterien für deren korrekte Bestimmung so klar, dass sie selbst von Studenten im ersten Semester nach kurzer Einführung korrekt anzuwenden sind. Gelingt es einem Tagger nicht – und bei den im Folgenden betrachteten Taggern[7] ist dies, wie zu zeigen sein wird, der Fall[8] – Präpositionen mit sehr hoher Wahrscheinlichkeit als solche zu erkennen, so ist davon auszugehen, dass sein Erkennungsverfahren zumindest unter diesem Aspekt (noch) nicht hinreichend weit entwickelt ist.

Da an eine umfassende Überprüfung der Verlässlichkeit der uns zur Verfügung stehenden Tagger selbst schon ein ebenso langwieriges wie aufwendiges Forschungsprojekt erforderlich machen müsste, konnten hier nur exemplarisch einige Auswertungen vorgenommen werden. Im Einzelnen sind dies Untersuchungen zu:

[7] Dabei handelt es sich um den bereits erwähnten TreeTagger und Connexor von Connexor Oy, Helsinki Business and Science Park, Finland. Beide wurden hier ausgewählt, weil es sich dabei um die Produkte handelt, die uns für unsere korpusgrammatischen Untersuchungen am Institut für Deutsche Sprache zur Verfügung stehen.

[8] Bei aller berechtigten Kritik ist Häme völlig unangebracht. Gerade eine Überprüfung der Akkuratesse dieser Tagger kann zeigen, welch enorme Leistung die Entwickler bereits erbracht haben, denn dabei wird schnell deutlich, wie überaus schwer es fallen muss, diese Aufgabe zu bewältigen.

- den – zugegebenermaßen äußerst problematischen – Wortformen *meine/Meine*;
- Präpositionen im Allgemeinen;
- den Wortformen *bar*, *eingangs*, *fernab*, *halber*, *jenseits*, *laut*.

### 4.2.1 Die Wortformen *meine/Meine*

**Laut GDS**

Auf der Basis der „Grammatik der deutschen Sprache" (GDS) kann man den Wortformen *meine/Meine* die folgenden Kategorien zuordnen und bei einer syntaktischen Analyse beliebiger deutscher Sätze auch ohne größere Schwierigkeiten erkennen, welche davon jeweils zutrifft:

- Verb: 1. Person Singular Indikativ Präsens
- Verb: 1. Person Singular Konjunktiv Präsens
- Verb: 3. Person Singular Konjunktiv Präsens
- Verb: Imperativ Singular
- Possessivartikel: 1. Person Singular Femininum Nominativ
- Possessivartikel: 1. Person Singular Femininum Akkusativ
- Possessivartikel: 1. Person Plural Mask./Fem./Neutr. Nominativ
- Possessivartikel: 1. Person Plural Mask./Fem./Neutr. Akkusativ
- Maskulines Nomenkonvertat
- Feminines Nomenkonvertat
- Neutrales Nomenkonvertat
- Eigenname

**Laut TreeTagger**

Der TreeTagger differenziert anders als GDS und, wie sich gleich zeigen wird, als Connexor:

- <Meine/NE> (Eigenname)
- <Meine/PPOSAT> (attribuierendes Possessivpronomen)
- <Meine/PPOSS> (substituierendes Possessivpronomen)
- <Meine/VVFIN> (finites Vollverb)
- <meine/PPOSAT> (attribuierendes Possessivpronomen)

- <meine/PPOSS> (substituierendes Possessivpronomen)
- <meine/VVFIN> (finites Vollverb)

Eine Testsuche mit Hilfe des Recherchesystems COSMAS II ergab[9]:

- Unter 285 150 Fundstellen für die Wortformen *Meine/meine* wurden 2 172 korrekt als Verb erkannt, doch immerhin 420 *false positives*.
- Als Nomen wird die Wortform *Meine* 87-mal gefunden. Darunter finden sich neun *false positives*. Hier einige Beispiele:

  ***Meine drei Kinder** sind aus dem Alter raus, in dem sie nach Ostereiern suchen.* [Berliner Ztg., 30.03.2002 (S. 23)]

  ***Meine Wirbelsäule** schmerzt, nach dreihundert Metern Gehstrecke habe ich zudem starke Schmerzen in beiden Beinen.* [Berliner Ztg., 10.02.2005 (S. 27)]

  ***Meine Reisetasche** ist schon gepackt.* [Berliner Ztg., 23.06.2005 (S. 22)]

- 78-mal wurde der Stadtname *Meine* gefunden, jedoch keineswegs als Eigenname, obwohl diese Kategorie grundsätzlich vorgesehen ist. Als Eigenname wurde die Wortform nie gefunden, was einigermaßen überrascht, da allein schon der Name des Leadsängers der Musikgruppe Scorpions – *Klaus Meine* – 250 mal vertreten ist, womit zugleich mindestens 250 *false negatives* zu verzeichnen sind.
- Die Auszeichnung von 48 098 Wortformen als attribuierende Possessivpronomina ist weitestgehend korrekt, doch ist hier mit zahlreichen *false negatives* zu rechnen. So fanden sich bei einer Suche mittels COSMAS unter 2 172 Fundstellen für *meine* als finite Verbform 114 „Treffer", bei denen es sich tatsächlich um Possessivpronomina handelte.
- Bei einer Suche nach *Meine/meine* **nicht** als Verbform fanden sich in einer Auswahl von 50 000 Treffern mindestens 1 910 Wortformen, die als Verbformen hätten erkannt werden müssen. Das sind zwar gerade mal knapp 4% *false positives*, doch zugleich extrem viele *false negatives* an Verbformen, wenn man bedenkt, dass nur 1 752 Wortformen korrekt als Verbformen erkannt wurden, denen – hochgerechnet – 11 319 nicht erkannte gegenüber stehen.

[9] Generell ist anzumerken, dass die Ergebnisse von Annotationen mit TreeTagger davon abhängen, welches Trainingskorpus zugrunde gelegt wurde. Bei den im Folgenden vorgenommenen Recherchen zu Präpositionen kann die Wahl eines geeigneten oder eher ungeeigneten Trainingskorpus entscheidenden Einfluss auf die Qualität der Ergebnisse gehabt haben. Bei so häufig verwendeten Wortformen wie *Meine* und *meine* sollte dies jedoch kaum eine Rolle spielen.

Das heißt: Selbst wenn die Suche nahezu nur korrekte Ergebnisse liefert, kann nicht in jeder Hinsicht davon ausgegangen werden, dass sie erfolgreich war.

**Laut Connexor Machinese Language Model**

Das Connexor Machinese Language Model unterscheidet bei den Wortformen *Meine/meine* diese Kategorien:

- <N > (Noun)
- <PRON> (Pronoun)
- <V IMP> (Verb, Imperative)
- <V IND PRES> (Verb, Indicative Present)
- <V SUB PRES> (Verb, Subjunctive Present)

Die vorgenommenen Unterscheidungen sind zwar in mancher Hinsicht etwas weniger differenziert als diejenigen der GDS, in anderer Hinsicht sogar weitergehend. Generell kann festgehalten werden, dass die Klassifikation auch für traditionell ausgerichtete Grammatiker grundsätzlich brauchbar sein sollte. Alles hängt jetzt davon ab, wie zutreffend die Annotation im Praxistest ausfällt. Hier einige Ergebnisse:

- Unter insgesamt 22 874 Fundstellen[10] für *meine/Meine* als Verbform waren 10644 fälschlich als solche ausgezeichnet. Man spricht in diesem Fall von *false positives*.
- Unter 67 Fundstellen für *meine/Meine* als Nomen waren 11 falsch ausgezeichnet.
- Unter 30 000 Fundstellen[11] für *meine* als Pronomen erwiesen sich zwar nur etwa drei Prozent als Fehleinschätzung, doch in die Bewertung muss auch eingehen, dass sich schon allein unter den 10 644 fälschlich als Verbform ausgezeichneten Fundstellen 7 804 Fundstellen befanden, bei denen *meine* als Pronomen hätte erkannt werden müssen. In diesem Fall spricht man von *false negatives*.

Insgesamt erwies sich Connexor bei diesen Tests als ebenso wenig leistungsfähig wie TreeTagger.

[10] Alle Suchen fanden im Zeitraum Juni bis Oktober 2011 mit COSMAS II, dem Recherchesystem des Instituts für Deutsche Sprache in Mannheim statt. Als Textbasis dienten die dort eingestellten, mit Connexor und TreeTagger annotierten Korpora.

[11] Gefunden bei interner Recherche am Institut für Deutsche Sprache in Mannheim.

### 4.2.2 Präpositionen

**Präpositionen laut *grammis*[12]**

Präpositionen sind in allen Grammatiken zum Deutschen als Wortart vertreten. Hier, wie in *grammis* – dem auf der GDS basierenden grammatischen Informationssystem des Instituts für Deutsche Sprache – Präpositionen eingeführt werden:

Präpositionen (z.B. *an*, *aus*, *von*, *zu*, *wider*, *entsprechend*, *ungeachtet*) sind unflektierbare Ausdrücke, die Gegenstände oder Sachverhalte in eine spezifische inhaltliche Beziehung zueinander setzen, z.B. in eine räumliche (*die Katze **auf** dem heißen Blechdach*), in eine kausale (*zitternd vor Angst*) oder in eine zeitliche (*Tod **um** Mitternacht*). Sie können einerseits vor oder – viel seltener – nach einer Nominalphrase bzw. Pronominalphrase (***für** die Kinder*, ***für** sie*, *den Kindern **zuliebe***, *ihnen **zuliebe***), andererseits vor einer Adverbphrase (***bis** morgen*) stehen. Mit diesen Phrasen zusammen bilden Präpositionen sogenannte Präpositionalphrasen. Dabei regieren sie den Kasus des Nomens bzw. Pronomens.[13]

Auf der Basis dieser Charakterisierung werden in der grammis-Komponente „Grammatisches Wörterbuch" die Wörter aus Tabelle 1 als Präpositionen geführt.[14]

Die Liste stützt sich ausschließlich auf „intelligente" Auswertungen von Texten sowie einschlägige Literatur. Unberücksichtigt bleiben in dieser Liste Abkürzungen wie etwa *wg.* für *wegen*, sowie Kombinationen aus Präposition und Artikel wie etwa *beim*, *vors*, *zur*.

[12] http://hypermedia.ids-mannheim.de/grammis/ (Stand: August 2013).

[13] http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht?v_typ=d&v_id=210 (Stand: August 2013).

[14] http://hypermedia.ids-mannheim.de/pls/public/gramwb.ansicht?v_app=g&v_kat=Pr%E4position&v_buchstabe= (Stand: August 2013).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1. | å | 34. | eingedenk | 67. | laut | 100. | südlich |
| 2. | ab | 35. | einschließlich | 68. | ledig | 101. | südöstlich |
| 3. | abseits | 36. | entgegen | 69. | links | 102. | südwestlich |
| 4. | abzüglich | 37. | entlang | 70. | mangels | 103. | trotz |
| 5. | an | 38. | entsprechend | 71. | minus | 104. | über |
| 6. | angesichts | 39. | exklusive | 72. | mit | 105. | um |
| 7. | anhand | 40. | fern | 73. | mithilfe | 106. | um ... willen |
| 8. | anlässlich | 41. | fernab | 74. | mitsamt | 107. | unbeschadet |
| 9. | anstatt | 42. | frei | 75. | mittels | 108. | unfern |
| 10. | anstelle | 43. | für | 76. | nach | 109. | ungeachtet |
| 11. | antwortlich | 44. | gegen | 77. | nächst | 110. | unter |
| 12. | auf | 45. | gegenüber | 78. | nah(e) | 111. | unterhalb |
| 13. | auf Grund | 46. | gelegentlich | 79. | namens | 112. | unweit |
| 14. | aus | 47. | gemäß | 80. | neben | 113. | vermittels |
| 15. | außer | 48. | getreu | 81. | nebst | 114. | vermöge |
| 16. | außerhalb | 49. | gleich | 82. | nördlich | 115. | via |
| 17. | ausgangs | 50. | halber | 83. | nordöstlich | 116. | vis å vis |
| 18. | ausgenommen | 51. | hinsichtlich | 84. | nordwestlich | 117. | voll |
| 19. | ausschließlich | 52. | hinter | 85. | ob | 118. | von |
| 20. | ausweislich | 53. | in | 86. | oberhalb | 119. | vor |
| 21. | bar | 54. | in puncto | 87. | ohne | 120. | vorbehaltlich |
| 22. | behufs | 55. | in Sachen | 88. | per | 121. | während |
| 23. | bei | 56. | infolge | 89. | plus | 122. | wegen |
| 24. | beiderseits | 57. | inklusive | 90. | pro | 123. | wider |
| 25. | betreffs | 58. | inmitten | 91. | rechts | 124. | zeit |
| 26. | bezüglich | 59. | innerhalb | 92. | rücksichtlich | 125. | zu |
| 27. | binnen | 60. | je | 93. | samt | 126. | zufolge |
| 28. | bis | 61. | jenseits | 94. | seit | 127. | zugunsten |
| 29. | dank | 62. | kontra | 95. | seitab | 128. | zuliebe |
| 30. | diesseits | 63. | kraft | 96. | seitens | 129. | zuungunsten |
| 31. | durch | 64. | lang | 97. | seitlich | 130. | zuzüglich |
| 32. | einbezüglich | 65. | längs | 98. | seitwärts | 131. | zwecks |
| 33. | eingangs | 66. | längsseits | 99. | statt | 132. | zwischen |

**Tab. 1:** Präpositionen laut *grammis*

## Präpositionen laut Connexor

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1. | "a" | 34. | "bis" | 67. | "mangels" | 100. | "ungeachtet" |
| 2. | "a****" | 35. | "bis"- | 68. | "mit" | 101. | "unter" |
| 3. | "ab" | 36. | "contra" | 69. | "mit"- | 102. | "unterhalb" |
| 4. | "ab"- | 37. | "contra"- | 70. | "mithilfe" | 103. | "unweit" |
| 5. | "abseits" | 38. | "dank" | 71. | "mittels" | 104. | "versus" |
| 6. | "abzüglich" | 39. | "diesseits" | 72. | "nach" | 105. | "versus"- |
| 7. | "ad" | 40. | "durch" | 73. | "nach"- | 106. | "via" |
| 8. | "ad"- | 41. | "entgegen" | 74. | "nah" | 107. | "via"- |
| 9. | "ähnlich" | 42. | "entlang" | 75. | "nahe" | 108. | "von" |
| 10. | "als" | 43. | "entsprechend" | 76. | "neben" | 109. | "von"- |
| 11. | "an" | 44. | "fern" | 77. | "neben"- | 110. | "vor" |
| 12. | "an"- | 45. | "für" | 78. | "nebst" | 111. | "vorbehaltlich" |
| 13. | "angesichts" | 46. | "für"- | 79. | "nördlich" | 112. | "während" |
| 14. | "anlässlich" | 47. | "gegen" | 80. | "ob" | 113. | "wegen" |
| 15. | "anläßlich" | 48. | "gegen"- | 81. | "ob"- | 114. | "wider" |
| 16. | "anno" | 49. | "gegenüber" | 82. | "oberhalb" | 115. | "wie" |
| 17. | "anno"- | 50. | "gen" | 83. | "ohne" | 116. | "wie"- |
| 18. | "anstatt" | 51. | "hinter" | 84. | "ohne"- | 117. | "zeit" |
| 19. | "anstatt"- | 52. | "in" | 85. | "per" | 118. | "zeit"- |
| 20. | "anstelle" | 53. | "in"- | 86. | "per"- | 119. | "zu" |
| 21. | "auf" | 54. | "in", | 87. | "plus" | 120. | "zu"- |
| 22. | "aus" | 55. | "in". | 88. | "pro" | 121. | "zufolge" |
| 23. | "ausschliesslich" | 56. | "infolge" | 89. | "rechts" | 122. | "zulasten" |
| 24. | "ausschließlich" | 57. | "inklusive" | 90. | "seit" | 123. | "zuliebe" |
| 25. | "außer" | 58. | "inmitten" | 91. | "seitens" | 124. | "zunächst" |
| 26. | "ausser"- | 59. | "innerhalb" | 92. | "sonder" | 125. | "zuzüglich" |
| 27. | "ausserhalb" | 60. | "je" | 93. | "sonder"- | 126. | "zwecks" |
| 28. | "außerhalb" | 61. | "jenseits" | 94. | "statt" | 127. | "zwischen" |
| 29. | "ausweislich" | 62. | "kontra" | 95. | "statt"- | 128. | "zwischen"- |
| 30. | "behufs" | 63. | "kraft" | 96. | "südlich" | 129. | a |
| 31. | "bei" | 64. | "längsseits" | 97. | "trotz" | 130. | å |
| 32. | "binnen" | 65. | "laut" | 98. | "über" | 131. | å- |
| 33. | "binnen"- | 66. | "links" | 99. | "um" | 132. | a) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 133. | a), | 166. | an-kopf-rennen | 199. | außer. | 232. | einschliesslich |
| 134. | a, | 167. | anlaesslich | 200. | ausserhalb | 233. | einschließlich |
| 135. | a. | 168. | anlässlich | 201. | außerhalb | 234. | entgegen |
| 136. | a.: | 169. | anläßlich | 202. | ausweislich | 235. | entgegen- |
| 137. | a» | 170. | anno | 203. | behufs | 236. | entlang |
| 138. | a», | 171. | anno- | 204. | bei | 237. | entlang- |
| 139. | ab | 172. | ans | 205. | bei- | 238. | entlang, |
| 140. | ab- | 173. | anstatt | 206. | bei, | 239. | entlang. |
| 141. | ab. | 174. | anstatt- | 207. | bei. | 240. | entsprechend |
| 142. | abseits | 175. | anstelle | 208. | bei-den | 241. | fern |
| 143. | abseits- | 176. | antwortlich | 209. | beidseits | 242. | fern. |
| 144. | ab-werkstätte | 177. | antwortlich. | 210. | beidseits- | 243. | fuer |
| 145. | abzüglich | 178. | auf | 211. | beim | 244. | fuer. |
| 146. | ad | 179. | auf- | 212. | betreffs | 245. | für |
| 147. | ad- | 180. | auf, | 213. | beziehentlich | 246. | für, |
| 148. | ad. | 181. | auf. | 214. | bezueglich | 247. | für. |
| 149. | aehnlich | 182. | auf» | 215. | bezüglich | 248. | fürs |
| 150. | ähnlich | 183. | auf», | 216. | binnen | 249. | gegen |
| 151. | ähnlich. | 184. | aufgrund | 217. | binnen. | 250. | gegen- |
| 152. | als | 185. | auf-hand-artistik | 218. | bis | 251. | gegen. |
| 153. | als- | 186. | aufs | 219. | bis- | 252. | gegenüber |
| 154. | als. | 187. | aus | 220. | bis. | 253. | gegenueber |
| 155. | als-ob. | 188. | aus- | 221. | contra | 254. | gemaess |
| 156. | am | 189. | aus. | 222. | contra- | 255. | gemäss |
| 157. | an | 190. | ausgangs | 223. | contra-referent | 256. | gemäß |
| 158. | an- | 191. | ausgangs- | 224. | dank | 257. | gemäss, |
| 159. | an. | 192. | ausschliesslich | 225. | dank- | 258. | gen |
| 160. | an-bug-rennen | 193. | ausschließlich | 226. | diesseits | 259. | gen- |
| 161. | anfangs | 194. | ausser | 227. | durch | 260. | gen. |
| 162. | anfangs- | 195. | außer | 228. | durch- | 261. | halber |
| 163. | angesichts | 196. | ausser- | 229. | durch. | 262. | hinsichtlich |
| 164. | anhand | 197. | außer- | 230. | durchs | 263. | hinter |
| 165. | an-kopf- | 198. | ausser. | 231. | eingangs | 264. | hinters |

| 265. | i | 298. | links | 331. | nördlich | 364. | seitlich |
|---|---|---|---|---|---|---|---|
| 266. | im | 299. | links- | 332. | nördlich- | 365. | seitlich- |
| 267. | im, | 300. | links. | 333. | nördlich. | 366. | sonder |
| 268. | in | 301. | mangels | 334. | nordoestlich | 367. | sonder, |
| 269. | in- | 302. | minus | 335. | nordöstlich | 368. | sonder. |
| 270. | in!» | 303. | minus- | 336. | ob | 369. | statt |
| 271. | in, | 304. | minus. | 337. | ob- | 370. | statt- |
| 272. | in. | 305. | mit | 338. | ob. | 371. | statt, |
| 273. | in» | 306. | mit- | 339. | oberhalb | 372. | statt. |
| 274. | in-box | 307. | mit. | 340. | ohne | 373. | südlich |
| 275. | infolge | 308. | mithilfe | 341. | ohne- | 374. | südlich- |
| 276. | in-formiert | 309. | mithilfe, | 342. | ohne. | 375. | suedlich |
| 277. | in-hand-arbeit | 310. | mithilfe. | 343. | ohne» | 376. | trotz |
| 278. | inklusive | 311. | mitsamt | 344. | per | 377. | trotz- |
| 279. | in-luft- | 312. | mittels | 345. | per- | 378. | über |
| 280. | inmitten | 313. | mit-treff | 346. | per gynt | 379. | über- |
| 281. | innerhalb | 314. | mit-treff» | 347. | per. | 380. | übers |
| 282. | ins | 315. | nach | 348. | per/pd | 381. | ueber |
| 283. | in-und | 316. | nach- | 349. | plus | 382. | um |
| 284. | je | 317. | nach. | 350. | plus- | 383. | um- |
| 285. | jenseits | 318. | nächst | 351. | plus. | 384. | um. |
| 286. | jenseits- | 319. | naechst | 352. | pro | 385. | ums |
| 287. | kontra | 320. | nah | 353. | pro. | 386. | unangesehen |
| 288. | kraft | 321. | nah. | 354. | punkto | 387. | unbeschadet |
| 289. | kraft- | 322. | nahe | 355. | punkto- | 388. | unbeschadet. |
| 290. | kraft. | 323. | nahe- | 356. | rechts | 389. | unerachtet |
| 291. | la | 324. | nahe. | 357. | rechts- | 390. | unfern |
| 292. | la- | 325. | namens | 358. | rücksichtlich | 391. | ungeachtet |
| 293. | längs | 326. | namens. | 359. | samt | 392. | unter |
| 294. | längsseits | 327. | neben | 360. | samt- | 393. | unter- |
| 295. | längsseits. | 328. | neben- | 361. | seit | 394. | unterhalb |
| 296. | laut | 329. | nebst | 362. | seit- | 395. | unterm |
| 297. | laut. | 330. | noerdlich | 363. | seitens | 396. | unters |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 397. | unweit | 414. | vor. | 431. | willen. | 448. | zuliebe», |
| 398. | vermittels | 415. | vorbehaltlich | 432. | zeit | 449. | zum |
| 399. | vermittelst | 416. | vors | 433. | zeit- | 450. | zu-mund-propaganda. |
| 400. | vermöge | 417. | waehrend | 434. | zeit, | 451. | zunächst |
| 401. | versus | 418. | während | 435. | zeit. | 452. | zunaechst |
| 402. | versus- | 419. | während- | 436. | zeit» | 453. | zur |
| 403. | via | 420. | wegen | 437. | zu | 454. | zur- |
| 404. | via- | 421. | wegen- | 438. | zu- | 455. | zu-span-zeiten |
| 405. | via. | 422. | wegen, | 439. | zu. | 456. | zuungunsten |
| 406. | vo | 423. | wegen. | 440. | zufolge | 457. | zuwider |
| 407. | vom | 424. | wegen.» | 441. | zufolge, | 458. | zuzueglich |
| 408. | von | 425. | wider | 442. | zufolge. | 459. | zuzüglich |
| 409. | von- | 426. | wider. | 443. | zugunsten | 460. | zwecks |
| 410. | von. | 427. | wie | 444. | zulasten | 461. | zwecks- |
| 411. | von-rohr | 428. | wie- | 445. | zulieb | 462. | zwischen |
| 412. | vor | 429. | wie-ein-pirat-tag | 446. | zuliebe | 463. | zwischen- |
| 413. | vor- | 430. | willen | 447. | zuliebe. | | |

**Tab. 2: Präpositionen laut Connexor**

Die Bestimmung der von Connexor als Präpositionen ausgezeichneten Wortformen erfolgte auf der Basis der damit annotierten DeReKo-Texte. Es könnte sich mithin nur um eine Auswahl handeln, doch in Anbetracht der Größe des Korpus dürfte nur wenig und Seltenes entgangen sein. Auffällig – und damit vielleicht ein Hinweis auf ernsthafte Schwierigkeiten – ist allerdings, dass diese drei Präpositionen in der Liste fehlen: *südöstlich*, *südwestlich*, *nordwestlich*, denn in den nicht-annotierten Korpora des DeReKo finden sich für diese Wortformen über 20.000 Belege.

Man erkennt schnell, dass Connexor offenbar Probleme mit der Erfassung und der Tokenisierung der genauen Wortformen hatte, was ganz oder teilweise auf die Kodierung der Ausgangsdaten zurückzuführen sein könnte. Tilgt man alles, was solchen Problemen geschuldet sein könnte, verbleiben ganze 159 Wortformen. Reduziert man die Liste weiter um Schreibvarianten – die zu erkennen durchaus richtig war – bleiben 149 Wortformen. Stellt man weiterhin in Rechnung, dass – anders als bei *grammis* – auch 15 Kombinationen von Präpositionen und Artikel (*am, ans, aufs, beim, durchs, fürs, hinters, ins, übers,*

*ums, unters, vom, vors, zum, zur*) berücksichtigt wurden, verbleibt eine Liste nahezu gleichen Umfangs wie bei *grammis*.

Doch hier ist Vorsicht geboten: Nachfolgende Punkte, Kommata, Bindestriche und Anführungszeichen nach vermeintlichen Präpositionen können Indizien für fehlerhafte Tokenisierung sein. So dürften etwa Punkte allenfalls nach Postpositionen auftreten, die Connexor – wie auch *grammis* und viele Grammatiken, jedoch nicht TreeTagger! – zu den Präpositionen zählt. Es bleibt also zu prüfen, ob tatsächlich Verwendungen als Postpositionen vorlagen. Dabei zeigt bereits eine grobe Überprüfung, dass es sich zumindest bei 64 der mit einem Punkt abgeschlossenen, als Präpositionen klassifizierten Wortformen nicht um Postpositionen handeln kann und somit eine falsche Annotation vorliegen muss.

**Präpositionen laut TreeTagger**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1. | a. | 20. | angesichts | 39. | ausser | 58. | dank. |
| 2. | a.: | 21. | anhand | 40. | ausser, | 59. | dank: |
| 3. | a.b | 22. | anlässlich | 41. | ausserhalb | 60. | dank», |
| 4. | a.ch- | 23. | ans | 42. | ausserhalb. | 61. | diesseits |
| 5. | a.ch-verbindung | 24. | anstatt | 43. | b. | 62. | durch |
| 6. | a.fagetti@ tagblatt.ch | 25. | anstelle | 44. | b.knellwolf@ tagblatt.ch | 63. | durch» |
| 7. | a.husistein@ bluewin.ch | 26. | auf | 45. | bei | 64. | durch», |
| 8. | a.k | 27. | auf&; | 46. | bei» | 65. | durchs |
| 9. | a.m.). | 28. | auf), | 47. | bei», | 66. | eingangs |
| 10. | a.nagel@ tagblatt.ch | 29. | auf» | 48. | beiderseits | 67. | eingangs, |
| 11. | a.rh. | 30. | auf», | 49. | beiderseits. | 68. | einschliesslich |
| 12. | ab | 31. | auf». | 50. | beidseits | 69. | entgegen |
| 13. | ab, | 32. | auf»: | 51. | beim | 70. | entlang |
| 14. | ab» | 33. | aufgrund | 52. | bezüglich | 71. | entlang, |
| 15. | ab», | 34. | aufs | 53. | binnen | 72. | entlang. |
| 16. | abzüglich | 35. | aufseiten | 54. | bis | 73. | exklusive |
| 17. | am | 36. | aus | 55. | dank | 74. | f.kurth@ tagblatt.ch |
| 18. | am» | 37. | ausgangs | 56. | dank! | 75. | fern |
| 19. | an | 38. | ausschliesslich | 57. | dank, | 76. | fern». |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 77. | für | 105. | in, | 133. | m.rivas@ bluewin.ch | 161. | o.. |
| 78. | für, | 106. | in: | 134. | m.scherrer@ tagblatt.ch | 162. | o.ä.) |
| 79. | für: | 107. | in» | 135. | m.tel.ch | 163. | o.-durchgang, |
| 80. | fürs | 108. | infolge | 136. | mangels | 164. | o.-vergleich |
| 81. | gegen | 109. | inklusive | 137. | minus | 165. | oberhalb |
| 82. | gegen, | 110. | inklusive) | 138. | mit | 166. | ohne |
| 83. | gegenüber | 111. | inklusive. | 139. | mit: | 167. | ohne. |
| 84. | gemäss | 112. | inmitten | 140. | mit» | 168. | ohne» |
| 85. | gen | 113. | innerhalb | 141. | mit», | 169. | ohne», |
| 86. | gleich | 114. | ins | 142. | mithilfe | 170. | östlich |
| 87. | halber | 115. | je | 143. | mitsamt | 171. | p.landmark@ tagblatt.ch |
| 88. | halber, | 116. | jenseits | 144. | mittels | 172. | p.loher@tagblatt.ch |
| 89. | her | 117. | kontra | 145. | n. | 173. | p.m» |
| 90. | herum | 118. | kraft | 146. | n.v | 174. | p.surber@ tagblatt.ch |
| 91. | hin | 119. | längs | 147. | nach | 175. | per |
| 92. | hinsichtlich | 120. | längsseits. | 148. | nach» | 176. | pro |
| 93. | hinter | 121. | laut | 149. | nach», | 177. | pro&; |
| 94. | hinterm | 122. | laut&; | 150. | nach». | 178. | pro), |
| 95. | hinters | 123. | m. | 151. | nah | 179. | samt |
| 96. | i. | 124. | m.). | 152. | nahe | 180. | seit |
| 97. | i.a., | 125. | m.b.egli@ bluewin.ch | 153. | namens | 181. | seit' |
| 98. | i.di. | 126. | m.b.h., | 154. | neben | 182. | seiten |
| 99. | i.honsel@ theatersg.ch | 127. | m.baer@nv-c.ch | 155. | nebst | 183. | seitens |
| 100. | i.p. | 128. | m.boehringer @sbw.edu | 156. | nördlich | 184. | seitlich |
| 101. | im | 129. | m.elsener@ tagblatt.ch | 157. | nordöstlich | 185. | statt |
| 102. | im, | 130. | m.giger@toggen-burgmedien.ch | 158. | nordwest-lich | 186. | südlich |
| 103. | in | 131. | m.jocham@ gmx.ch | 159. | o. | 187. | südöstlich |
| 104. | in!» | 132. | m.loeliger@ tagblatt.ch | 160. | o., | 188. | trotz |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 189. | u. | 206. | unters | 223. | vorm | 240. | zeit |
| 190. | u.a. | 207. | unweit | 224. | vors | 241. | zu |
| 191. | über | 208. | v. | 225. | während | 242. | zu» |
| 192. | überm | 209. | v.chr. | 226. | wegen | 243. | zu», |
| 193. | übers | 210. | vermöge | 227. | wegen, | 244. | zufolge |
| 194. | um | 211. | versus | 228. | wegen. | 245. | zugunsten |
| 195. | um», | 212. | via | 229. | westlich | 246. | zuliebe |
| 196. | ums | 213. | vom | 230. | wider | 247. | zuliebe, |
| 197. | ungeachtet | 214. | vom, | 231. | willen | 248. | zuliebe. |
| 198. | unter | 215. | von | 232. | willen! | 249. | zum |
| 199. | unter!» | 216. | von, | 233. | willen, | 250. | zur |
| 200. | unter, | 217. | von. | 234. | willen. | 251. | zuungunsten |
| 201. | unter. | 218. | von: | 235. | willen: | 252. | zuzüglich |
| 202. | unter: | 219. | von« | 236. | willen? | 253. | zwecks |
| 203. | unterhalb | 220. | vor | 237. | willen» | 254. | zwischen |
| 204. | unterm | 221. | vor. | 238. | z. | | |
| 205. | untern | 222. | vorbehaltlich | 239. | z.b. | | |

**Tab. 3: Präpositionen laut TreeTagger**

Die Liste wurde auf der Basis fünfmaliger Zufallsauswahl von jeweils 100 000 Treffern aus insgesamt 153 982 558 unter den mit TreeTagger annotierten Texten des DeReKo erstellt und könnte deshalb – wie die Liste zu Connexor – unvollständig sein, doch dass viele wesentliche Elemente fehlen ist wenig wahrscheinlich. Bemerkenswert ist, dass in dieser Liste von den bei Connexor vermissten Präpositionen nur *südwestlich* fehlt, was hier tatsächlich der Zufallsauswahl geschuldet sein könnte.

Offensichtlich ist hingegen auch hier, dass die Liste zahlreiche Ausdrucksformen enthält, die so weder als Präpositionen noch überhaupt als Wortformen gelten können. Reduziert man die Liste um diese Elemente, verbleiben noch 129 Wortformen, unter denen sich 18 Kombinationen von Präposition mit Artikel befinden (*am, ans, aufs, beim, durchs, hinterm, hinters, im, ins, ums, unterm, untern, unters, vom, vorm, vors, zum, zur*).

Inwieweit die letztlich im Vergleich mit *grammis* und Connexor wesentlich geringere Zahl an erkannten Präpositionen auf die Unwägbarkeiten der Zufallsauswahl zurückzuführen ist, lässt sich mit den mir zur Verfügung stehen-

den Mitteln nicht überprüfen. Festzuhalten ist jedoch, dass die Suche nach Präpositionen zumindest auf der Basis von DeReKo nicht zu verlässlichen Daten führt, und da Präpositionen ein Bestandteil sehr vieler Phrasen sind, kann daraus geschlossen werden, dass auch darauf aufbauende Suchen nach komplexeren Strukturen wenig verlässlich sein werden.

### Kleine Zwischenbilanz

Was sich, bereinigt um Schreibungsvarianten, fälschlich als Wortformen gewerteten Ausdrücken sowie Präpositionen mit Artikel ergibt, zeigt Tabelle 4.

| *grammis* | | Connexor | | TreeTagger | |
|---|---|---|---|---|---|
| 1. | à | 1. | a | 1. | ab |
| 2. | ab | 2. | ab | 2. | abzüglich |
| 3. | abseits | 3. | abseits | 3. | an |
| 4. | abzüglich | 4. | abzüglich | 4. | angesichts |
| 5. | an | 5. | ad | 5. | anhand |
| 6. | angesichts | 6. | ähnlich | 6. | anlässlich |
| 7. | anhand | 7. | als | 7. | anstatt |
| 8. | anlässlich | 8. | an | 8. | anstelle |
| 9. | anstatt | 9. | anfangs | 9. | auf |
| 10. | anstelle | 10. | angesichts | 10. | aufgrund |
| 11. | antwortlich | 11. | anhand | 11. | aufseiten |
| 12. | auf | 12. | anläßlich | 12. | ausgangs |
| 13. | auf Grund | 13. | anno | 13. | ausschliesslich |
| 14. | aus | 14. | anstatt | 14. | ausser |
| 15. | außer | 15. | anstelle | 15. | ausserhalb |
| 16. | außerhalb | 16. | antwortlich | 16. | bei |
| 17. | ausgangs | 17. | auf | 17. | beiderseits |
| 18. | ausgenommen | 18. | aufgrund | 18. | beidseits |
| 19. | ausschließlich | 19. | aus | 19. | bezüglich |
| 20. | ausweislich | 20. | ausgangs | 20. | binnen |
| 21. | bar | 21. | ausschließlich | 21. | bis |
| 22. | behufs | 22. | außer | 22. | dank |
| 23. | bei | 23. | außerhalb | 23. | diesseits |

| *grammis* | | Connexor | | TreeTagger | |
|---|---|---|---|---|---|
| 24. | beiderseits | 24. | ausweislich | 24. | durch |
| 25. | betreffs | 25. | behufs | 25. | eingangs |
| 26. | bezüglich | 26. | bei | 26. | einschliesslich |
| 27. | binnen | 27. | beidseits | 27. | entgegen |
| 28. | bis | 28. | betreffs | 28. | entlang |
| 29. | dank | 29. | beziehentlich | 29. | exklusive |
| 30. | diesseits | 30. | bezüglich | 30. | fern |
| 31. | durch | 31. | binnen | 31. | für |
| 32. | einbezüglich | 32. | bis | 32. | gegen |
| 33. | eingangs | 33. | contra | 33. | gegenüber |
| 34. | eingedenk | 34. | dank | 34. | gemäß |
| 35. | einschließlich | 35. | diesseits | 35. | gen |
| 36. | entgegen | 36. | durch | 36. | gleich |
| 37. | entlang | 37. | eingangs | 37. | halber |
| 38. | entsprechend | 38. | einschließlich | 38. | her |
| 39. | exklusive | 39. | entgegen | 39. | herum |
| 40. | fern | 40. | entlang | 40. | hin |
| 41. | fernab | 41. | entsprechend | 41. | hinsichtlich |
| 42. | frei | 42. | fern | 42. | hinter |
| 43. | für | 43. | für | 43. | in |
| 44. | gegen | 44. | gegen | 44. | infolge |
| 45. | gegenüber | 45. | gegenüber | 45. | inklusive |
| 46. | gelegentlich | 46. | gemäß | 46. | inmitten |
| 47. | gemäß | 47. | gen | 47. | innerhalb |
| 48. | getreu | 48. | halber | 48. | je |
| 49. | gleich | 49. | hinsichtlich | 49. | jenseits |
| 50. | halber | 50. | hinter | 50. | kontra |
| 51. | hinsichtlich | 51. | in | 51. | kraft |
| 52. | hinter | 52. | infolge | 52. | längs |
| 53. | in | 53. | inklusive | 53. | laut |
| 54. | in puncto | 54. | inmitten | 54. | mangels |
| 55. | in Sachen | 55. | innerhalb | 55. | minus |

| *grammis* | | Connexor | | TreeTagger | |
|---|---|---|---|---|---|
| 56. | infolge | 56. | je | 56. | mit |
| 57. | inklusive | 57. | jenseits | 57. | mithilfe |
| 58. | inmitten | 58. | kontra | 58. | mitsamt |
| 59. | innerhalb | 59. | kraft | 59. | mittels |
| 60. | je | 60. | längs | 60. | nach |
| 61. | jenseits | 61. | längsseits | 61. | nah |
| 62. | kontra | 62. | laut | 62. | nahe |
| 63. | kraft | 63. | links | 63. | namens |
| 64. | lang | 64. | mangels | 64. | neben |
| 65. | längs | 65. | minus | 65. | nebst |
| 66. | längsseits | 66. | mit | 66. | nördlich |
| 67. | laut | 67. | mithilfe | 67. | nordöstlich |
| 68. | ledig | 68. | mitsamt | 68. | nordwestlich |
| 69. | links | 69. | mittels | 69. | oberhalb |
| 70. | mangels | 70. | nach | 70. | ohne |
| 71. | minus | 71. | nächst | 71. | östlich |
| 72. | mit | 72. | nah | 72. | per |
| 73. | mithilfe | 73. | nahe | 73. | pro |
| 74. | mitsamt | 74. | namens | 74. | samt |
| 75. | mittels | 75. | neben | 75. | seit |
| 76. | nach | 76. | nebst | 76. | seiten |
| 77. | nächst | 77. | nördlich | 77. | seitens |
| 78. | nah(e) | 78. | nordöstlich | 78. | seitlich |
| 79. | namens | 79. | ob | 79. | statt |
| 80. | neben | 80. | oberhalb | 80. | südlich |
| 81. | nebst | 81. | ohne | 81. | südöstlich |
| 82. | nördlich | 82. | per | 82. | trotz |
| 83. | nordöstlich | 83. | plus | 83. | über |
| 84. | nordwestlich | 84. | pro | 84. | um |
| 85. | ob | 85. | punkto | 85. | ungeachtet |
| 86. | oberhalb | 86. | rechts | 86. | unter |
| 87. | ohne | 87. | rücksichtlich | 87. | unterhalb |

| *grammis* | | Connexor | | TreeTagger | |
|---|---|---|---|---|---|
| 88. | per | 88. | samt | 88. | unweit |
| 89. | plus | 89. | seit | 89. | vermöge |
| 90. | pro | 90. | seitens | 90. | versus |
| 91. | rechts | 91. | seitlich | 91. | via |
| 92. | rücksichtlich | 92. | sonder | 92. | von |
| 93. | samt | 93. | statt | 93. | vor |
| 94. | seit | 94. | südlich | 94. | vorbehaltlich |
| 95. | seitab | 95. | trotz | 95. | während |
| 96. | seitens | 96. | über | 96. | wegen |
| 97. | seitlich | 97. | um | 97. | westlich |
| 98. | seitwärts | 98. | unangesehen | 98. | wider |
| 99. | statt | 99. | unbeschadet | 99. | willen |
| 100. | südlich | 100. | unerachtet | 100. | zeit |
| 101. | südöstlich | 101. | unfern | 101. | zu |
| 102. | südwestlich | 102. | ungeachtet | 102. | zufolge |
| 103. | trotz | 103. | unter | 103. | zugunsten |
| 104. | über | 104. | unterhalb | 104. | zuliebe |
| 105. | um | 105. | unweit | 105. | zuungunsten |
| 106. | um ... willen | 106. | vermittels | 106. | zuzüglich |
| 107. | unbeschadet | 107. | vermittelst | 107. | zwecks |
| 108. | unfern | 108. | vermöge | | |
| 109. | ungeachtet | 109. | versus | | |
| 110. | unter | 110. | via | | |
| 111. | unterhalb | 111. | von | | |
| 112. | unweit | 112. | vor | | |
| 113. | vermittels | 113. | vorbehaltlich | | |
| 114. | vermöge | 114. | während | | |
| 115. | via | 115. | wegen | | |
| 116. | vis à vis | 116. | wider | | |
| 117. | voll | 117. | wie | | |
| 118. | von | 118. | willen | | |
| 119. | vor | 119. | zeit | | |

| *grammis* | | Connexor | | TreeTagger | |
|---|---|---|---|---|---|
| 120. | vorbehaltlich | 120. | zu | | |
| 121. | während | 121. | zufolge | | |
| 122. | wegen | 122. | zugunsten | | |
| 123. | wider | 123. | zulasten | | |
| 124. | zeit | 124. | zulieb | | |
| 125. | zu | 125. | zuliebe | | |
| 126. | zufolge | 126. | zunächst | | |
| 127. | zugunsten | 127. | zuungunsten | | |
| 128. | zuliebe | 128. | zuwider | | |
| 129. | zuungunsten | 129. | zuzüglich | | |
| 130. | zuzüglich | 130. | zwecks | | |
| 131. | zwecks | 131. | zwischen | | |
| 132. | zwischen | | | | |

**Tab. 4: Vergleichstabelle**

Die Zahl der *false negatives* hält sich bei Connexor und TreeTagger in Grenzen, soweit es nur darum geht, Wortformen überhaupt als Präpositionen zu erkennen. An eindeutigen *false positives* finden sich nur zwei Wortformen bei TreeTagger: *hin* und *her*.

Die Schnittmenge der von beiden Taggern verzeichneten Präpositionen umfasst jedoch nur 98 Wortformen. Völlig offen bleibt so weit allerdings noch, wie gut die Erkennung von Fall zu Fall gelingt. Wie die im Folgenden beschriebenen exemplarischen Evaluationen zeigen werden, besteht in dieser Frage durchaus Anlass zu Bedenken.

### 4.2.3 Einzeluntersuchungen zur Erkennungsleistung

Die „Grammatik der deutschen Sprache“ und gleichermaßen *grammis* betrachten die in der untenstehenden Tabelle aufgeführten Lemmata als Präpositionen bzw. Postpositionen. TreeTagger und Connexor weisen diesen Lemmata die folgenden Annotationen zu:

| Lemma | TreeTagger | Connexor |
|---|---|---|
| *bar* | ADJD | A, N |
| *eingangs* | APPR, N | N, PREP |
| *fernab* | ADV | ADV, N |
| *halber* | ADJA, APPO | ADV, PREP |
| *jenseits* | ADV, APPR | ADV, N, PREP |
| *laut* | ADJD, APPR | A, N, PREP |

Tab. 5: Klassifikation verschiedener Tagger

Dass die Tagger in diesen Wortformen auch Elemente anderer Wortklassen erkennen können, beeinträchtigt ihre Brauchbarkeit nicht grundsätzlich, denn zum einen ist davon auszugehen, dass diese Wortformen auch andere Verwendungen als die einer Präposition haben können, zum anderen könnten die Tagger zwar andere Grundeinteilungen vornehmen als man selbst, jedoch mit etwas Flexibilität auch für die eigenen Zwecke nutzbar bleiben. Ob und, wenn ja, in wie weit dies der Fall ist, sollen die folgenden Einzelstudien zeigen.

## *bar*

### TreeTagger

Die Suche nach *bar* als beliebige Form der Präposition liefert keine Treffer, ebenso auch die Suche nach *bar* als Adverb, – ein Ergebnis, das im Hinblick auf die Annotationen bei vergleichbaren Wortformen von Bedeutung ist. Sucht man hingegen nach der Wortfolge *Bar/bar* jeder erhält man 1 959 Treffer. Dabei wird bemerkenswerterweise das großgeschriebene *Bar* 1 382-mal korrekt als Teil des Namens der Berliner Lokalität „Bar jeder Vernunft" als appelatives Nomen erkannt. Die restlichen 577 Treffer werden als prädikative Adjektive klassifiziert, eine Einschätzung die nicht unplausibel ist und sogar Anlass dafür bieten sollte, die Einschätzung in *grammis* zu überdenken, womit dann bereits ein Schritt in Richtung auf die geplante Evaluation von Aussagen zur deutschen Grammatik getan wäre.

Bleibt zu prüfen, wie TreeTagger die analogen Probleme etwa bei *fernab, jenseits* und *links* behandelt.

### Connexor

Auch Connexor liefert keine Treffer für *bar* als Präposition. Die generelle Suche nach *bar* als Wortform erbrachte am 18.11.2010 insgesamt 39 522 Treffer, unter denen 3 504 ganz ohne Auszeichnung blieben und 7 437 weitere nicht wirklich die Wortform *bar* auszeichneten, sondern *bar* + ein oder mehrere weitere Zeichen. Von den restlichen 28 581 Treffern waren 18 561 als Nomina ausgezeichnet, unter denen sich nur 17 falsche Auszeichnungen befanden. Bei weiteren 4 387 Treffern wurde großgeschriebenes *Bar* als Eigenname ausgezeichnet, das jedoch durchweg als Fehler betrachtet werden kann, weil nicht *Bar* selbst, sondern das jeweils vorangehende Nomen als Eigenname auszuzeichnen gewesen wäre, was aber nicht der Fall war.[15] Unter den 5 633 als Adjektiv ausgezeichneten Wortformen handelt es sich bei mehr als der Hälfte genau genommen um Adverbien, eine Kategorie, die Connexor zwar grundsätzlich kennt, jedoch nicht erkennt, wenn etwa von *bar bezahlen* die Rede ist. Bei den übrigen, als Adjektive ausgezeichneten Treffern unterscheidet Connexor nicht zwischen dem auf einen Zahlwert folgenden *bar*, das sich auf Druckverhältnisse bezieht und dem einwertigen *bar*, das man wie *grammis* als Präposition oder – besser – wie TreeTagger als prädikatives einwertiges Adjektiv bestimmen könnte.

Zugunsten von Connexor könnte man darauf hinweisen, dass eine große Zahl der Fehler auf die Art und Formatierung der Rohdaten zurückzuführen sein könnte, doch für Projekte, die darauf angewiesen sind, die Daten von DeReKo zu nutzen, ist dieser Hinweis wenig hilfreich.

## *eingangs*

### TreeTagger

Eine Suche nach den Wortformen *eingangs*/*Eingangs* als Präpositionen liefert 4 010 Treffer, doch schon ein kurzer Blick auf die Liste der *KWIC*[16] zeigt, dass dabei zahllose Fehleinschätzungen vorliegen. Hier einige Beispiele:

[15] So etwa in diesem Beleg: Die<DET @PREMOD> Firobig-<N @PREMOD>**<B>Bar</><N Prop @NH>** wird<V IND PRES @AUX> nicht<ADV @ADVL> mehr<ADV @ADVL> im<PREP @PREMARK> Sonnental<N @NH> eingerichtet,<V PCP PERF @MAIN> sondern<CC @CC> im<PREP @PREMARK> Barzelt<N @NH> beim<PREP @POSTMOD> Restaurant<N @PREMOD> Schäfli<N @NH> im<PREP @POSTMOD> Bernhardzeller<N @PREMOD> Schöntal.<N @NH>

[16] *KWIC* kurz für *Key Word In Context*

*Anschlussfehler, wie in der **eingangs** erwähnten Indy-Szene. Da sind*

*kommentiert der Trainer die **eingangs** geschriebene Entscheidung.*

*Hermann Bürgi erklärte **eingangs**, dass er nicht als*

*erweitert. Wie bereits **eingangs** ausgeführt, wurde am 2. Juni*

Eine – nicht bis ins letzte Detail durchgeführte – Überprüfung ergab über 2900 fehlerhafte Annotationen.

Eine Suche nach denselben Wortformen, jedoch nicht als Präpositionen ausgezeichnet, führt zu 1454 Treffern, unter denen sich neun Präpositionen befanden.

Insgesamt kann die Erkennungsleistung damit kaum als ausreichend betrachtet werden.

**Connexor**

Die Wortform *eingangs* wird 1693-mal als Präposition bestimmt, doch, wie eine – nicht bis ins letzte Detail durchgeführte – Überprüfung zeigt, befinden sich darunter mindestens[17] 870 falsche Einschätzungen, so etwa:

*So kommen bei einer Reihe von Autoimmunerkrankungen offensichtlich dieselben Effekte zum Tragen wie auch bei der **eingangs** genannten ENL-Therapie.* [taz, 03.02.2006 (S. 18)]

*In der **eingangs** erwähnten Stelle zur abendländischen Arroganz wiederum sagt Barthes in seiner Vorlesung, dass die Reinform der Arroganz die Ausbeutung der Evidenzen sei, die für selbstverständlich erklärt, was sie triumphieren lassen möchte.* [taz, 06.02.2006 (S. 15)]

*Wenn er eine Ballade trällert, leuchten auf sein Kommando alle **eingangs** verteilten Feuerzeuge auf und es wird klar, wo die absurden Kirmesneonarmbänder nach Dombesuchen so landen.* [taz, 21.02.2006 (S. 24)]

*Sorgenfrei der **eingangs** zitierte Springer-Reisende, der genüg mattes Gold dabei hat.* [taz, 04.03.2006 (S. 13)]

*Tja, und kaum dreißig Jahre nach dem **eingangs** geschilderten Zugerlebnis sitze ich in meiner Dachgeschoss-Maisonette vor dem eingebauten Kamin,*

[17] „Mindestens" soll heißen: Bei noch genauerer Überprüfung hätten sich vermutlich weitere falsche Einschätzungen finden lassen. Die Überprüfung wurde abgebrochen, weil das Ergebnis auch so schon aussagekräftig genug war.

*blättere abwechselnd in Sternes „Tristram Shandy" und Ulrich Greiners „Leseverführer", rülpse, streiche mir über meine seltsam rötliche Nase und lächle über die Sorge mancher alter Mitkämpfer, ihre mögliche Bürgerlichkeit oder Angepasstheit betreffend.* [taz, 20.03.2006 (S. 14)]

*Und womöglich ist „Surfaces" ja auch, gerade weil es sich ein wenig Humor leistet, ein Ausweg aus der* ***eingangs*** *erwähnten Krise.* [taz, 31.03.2006 (S. 28)]

Unter den 3 770 nicht als Präposition ausgezeichneten Wortformen fanden sich wiederum mindestens 310, die als Präpositionen zu erkennen gewesen wären. Insgesamt mithin auch hier eine eher unzureichende Erkennungsleistung.

### *fernab*

#### TreeTagger

Insgesamt 4 104, durchweg als Adverbien ausgezeichnet. Davon wären nach den Kriterien von *grammis* über 3 900 als Präpositionen auszuzeichnen. Einige Beispiele:

***Fernab*** *aller hektischen Betriebsamkeit läuft, auf gut 1500 Metern über Meer, die Fäneren zu ihrer Spitze aus.*

*Wer seine Erholung* ***fernab*** *vom städtischen Rummel sucht, der entsteigt dem City Night Line bereits in Kolding, einer Stadt, die im geografischen Zentrum von Dänemark an der Ostseeküste liegt.*

*Aber hier oben – fernab von blankgefegten und spuckefreien Trottoirs, wie sie sich manch ein Gossauer wohl wünscht – stört das niemanden.*

*Oder handelt es sich – so etwa Kritiker Hans Fässler – um die Identitätstümelei eines bestimmten Milieus,* ***fernab*** *der heutigen Realitäten?*

*Die Mitglieder des FCM würden auch* ***fernab*** *des Hafenfeld-Areals regelmässige Leistungen für die Gemeinschaft erbringen.*

Von fehlerhafter Auszeichnung kann hier insofern gesprochen werden, als nach den Kriterien, die im Fall von *bar* angewandt wurden, allenfalls eine Kategorisierung als prädikatives Adjektiv zu erwarten wäre.

**Connexor**

Insgesamt 4069 Treffer, keiner davon als Präposition ausgezeichnet. Nach *grammis* wären davon mindestens 3575 als Präpositionen einzustufen, wären also im Sinn der Recherche als *false negatives* einzuschätzen. Auch hier einige Beispiele:

*Das Trio ist mit seinen Kompositionen **fernab** des Wohlfühl-Jazz bestens für waghalsige Expeditionen gerüstet.* [St. Galler Tagbl., 27.10.2009 (S. 16)]

*Still und wie ein Gebet sang Zünd dieses weit, **fernab** allen grölenden Stolzes grosser Fussballfanmassen.* [St. Galler Tagbl., 03.03.2009 (S. 31)]

*Neben den für einmal in die Abendstunden verlegten Lektionen bot der erstmalig durchgeführte Anlass den Besucherinnen und Besuchern auch Attraktionen **fernab** des Schulbetriebs.* [St. Galler Tagbl., 21.03.2009 (S. 41)]

*Temperamentvoll und differenziert, aber **fernab** von schunkelseliger Volksmusik – auch wenn es ab und zu durchaus „heimelig" zu werden verspricht.* [St. Galler Tagbl., 25.03.2009 (S. 34)]

Alles in allem eine unbefriedigende Leistung.

***halber***

**TreeTagger**

Die Suche nach *halber* als Präposition (genauer: Postposition) führt zu 1764 Treffern. Davon sind 1687 korrekt als Postpositionen erkannt. Bei den restlichen 77 „Treffern" handelt es sich um attribute Adjektive und Adverbia. Mit unter 5% *false positives* fällt die Suche hier nicht schlecht aus. Da sich auch die Zahl der *false negatives* mit 65 unter den 4829 Treffern in engen Grenzen hält, kann hier von einer guten Erkennungsleistung gesprochen werden. Hier einige der wenigen verkannten Präpositionen:

*Der stellvertretende Redaktionsleiter von «SF Meteo», Christoph Siegrist, lebt **der Liebe halber** in St. Gallen.n* [St. Galler Tagbl., 30.10.2009 (S. 41)]

*Nennen wir ihn hier, **der Einfachheit halber**, also mal kurz „homo carnivorus". Ein Steak-Liebhaber, der seinen geliebten Rindern nur ungern das Grünzeug wegknabbert.* [Braunschw. Z., 11.08.2009]

*Klaus Volkert selbst aber kam dem Verfahren nun zuvor – und verzichtete bereits vor Haftantritt auf „Grad und Würde eines Doktors der Staatswissenschaften* ***Ehren halber***". [Braunschw. Z., 21.12.2009]

*Zuerst* ***der Sicherheit halber*** *die historischen Zusammenhänge: ...* [St. Galler Tagblatt, 08.11.2000]

**Connexor**

Unter den 1 436 Fundstellen finden sich 46 eindeutige Fehleinschätzungen, also gerade mal 3%. Unter den 5 138 ausdrücklich nicht als Präpositionen ausgezeichneten Fundstellen fanden sich 357, die als Präpositionen auszuzeichnen gewesen wären, also in etwa 6,9% *false negatives*. Insgesamt mithin eine achtbare Erkennungsleistung, die vielleicht darauf zurückzuführen sein könnte, dass es sich bei *halber* um eine Postposition handelt, der – anders als bei echten Präpositionen – stets ein Nomen unmittelbar voran gehen muss.

Einige Beispiele von *false negatives*:

*Auch hier sei* ***der Ordnung halber*** *noch angefügt: Nur drei Minuten zuvor stand Parhizi nach schönem Doppelpass mit De Simone alleine vor Schneider, doch dessen Geschoss entschärfte der Torsteher aus wenigen Metern mirakulös.* [St. Galler Tagbl., 06.04.2009 (S. 34)]

*Alle Tierhalter sollten* ***der Fairness halber*** *unabhängig vom Zeitpunkt der Impfung den Beitrag gemäss dem neuen Artikel in der Verordnung zahlen, findet der Regierungsrat.* [St. Galler Tagbl., 28.04.2009 (S. 25)]

*lange nämlich, bevor ich irgendeinen Namen dafür besaß oder mit auch nur von ihrer weiteren und allgemeinen Bedeutung ein Bild zu machen wußte, so daß ich die lebhafte Neigung zu gewissen Vorstellung und das durchdringende Vergnügen daran durch geraume Zeit für eine ganz persönliche und anderen gar nicht verständliche Eigentümlichkeit hielt, über* ***die ihrer Sonderbarkeit halber*** *lieber nicht zu sprechen sei.* [Th. Mann, Werke; Felix Krull, (1. Buchausg. 1954), 1960, Bd. 7 (S. 311)]

***jenseits***

**TreeTagger**

In einer Zufallsauswahl von 10 000 – aus insgesamt 21 193 – Fundstellen für *jenseits* als Präposition fanden sich lediglich 28 *false positives*. Dies ist bemerkenswert, nicht nur wegen der hohen Erkennungsleistung, sondern vor allem deshalb, weil *jenseits* offenbar anders eingeschätzt wird als *fernab*, zwei Wortformen, die nicht nur von *grammis* und der „Grammatik der deutschen Sprache" als von gleicher Art betrachtet werden. Gründe für die ungleiche Einschätzung durch TreeTagger sind für die Nutzer nicht erkennbar. Das hat unter anderem zur Folge, dass generelle Suchen nach Ausdrucksstrukturen, die von Präpositionen eingeleitet werden, nicht zu brauchbaren Ergebnissen führen können. Für ein Projekt, das Aussagen über syntaktische Strukturen im Deutschen auf der Basis von Textkorpora evaluieren will, können sich solche Inkongruenzen fatal auswirken.

Eine Überprüfung der 7 609 Treffer bei einer Suche nach den Wortformen *jenseits/Jenseits* die nicht als Präpositionen ausgezeichnet wurden, erbrachte mindestens 4 592 *false negatives*, also eine Fehlerquote von über 60%. Nicht als Präposition erkannt wurden *Jenseits/jenseits* unter anderem in diesen Sätzen:

> *Der gewohnte Anblick soll die Fans von der Insel und von* ***jenseits des Rheins*** *beruhigen.* [Zeit, 23.03.2006 (S. 12)]
>
> *An der Prager Filmhochschule entstehen Animationsfilme* ***jenseits von Disney****.* [Zeit, 30.03.2006 (S. 20)]
>
> *Mit dem Wachstumsschub bis zum Jahr 2011, der laut Frenkel auch die Schaffung weiterer Arbeitsplätze auf dem Vallendarer Campus nach sich ziehen wird, ist es aber für den Rektor der WHU nicht getan: „****Jenseits dieser mittelfristigen Strategie*** *muss das Ziel für uns 1000 bis 1500 Studenten heißen."* [RZ, 02.10.2009]
>
> *Ein hochrangiger Sicherheitsbeamter habe von Falschinformationen „****jenseits seiner Vorstellungen****" gesprochen.* [Berliner Ztg., 17.03.2004 (S. 6)]

### Connexor

In einer Auswahl von 10 000 Treffern – aus insgesamt 14 135 – fanden sich ganze neun *false positives* für *jenseits* als Präposition. Auch hier ist anzumerken, dass der Tagger *jenseits* anders einschätzt als das nach theoretischen Kriterien gleichartige *fernab*, was gravierende Folgen für generelle Suchen hat.

Sehr viel weniger erfolgreich stellt sich die Suche allerdings dar, wenn man betrachtet, was der Tagger alles nicht erkannt hat, obwohl er es nach eben den Kriterien, die zu den korrekten Funden führten, hätte erkennen müssen: Eine Überprüfung einer Auswahl von 10 000 Treffern bei einer Suche nach den Wortformen *jenseits/Jenseits* die nicht als Präpositionen ausgezeichnet wurden, erbrachte mindestens 6 167 *false negatives*, also eine Fehlerquote von über 60%. Einige Beispiele hiervon:

> ***Jenseits der vielen Rätsel*** *um den jährlich wiederkehrenden Streit um viel Geld gibt es zwischen Moskau und Kiew aber einen grundsätzlichem politischen Dissens.* [St. Galler Tagbl., 17.01.2009 (S. 7)]
>
> *Langfristig haben FDP und Grüne aber eine Alternative* ***jenseits des Links-Rechts-Schemas.*** [St. Galler Tagbl., 20.01.2009 (S. 7)]
>
> *Gönnen Sie sich eine Auszeit und nehmen Sie das Buch «Geld und Geist» von Jeremias Gotthelf und das bleibende Werk von Friedrich Walti «****Jenseits der Strasse****» mit.* [St. Galler Tagbl., 22.01.2009 (S. 27)]
>
> ***Jenseits des Klischees*** *vom „gemütlichen" Haydn, dessen Tod vor zweihundert Jahren man 2009 gedenkt, soll er als höchst unorthodoxer, innovativer Komponist aufscheinen.* [St. Galler Tagbl., 06.02.2009 (S. 45)]
>
> *Befürchtet „Pro Reli", dass die jungen Menschen sich eine eigene* ***Meinung jenseits vom Christentum oder überhaupt jenseits von Religionen*** *bilden?* [Rheinpfalz, 27.12.2008, S. 60]

## *laut*

### TreeTagger

9 936 korrekte Fundstellen unter 10 000 zufällig aus 398 500 Treffern ausgewählten Belegen können als hervorragendes Ergebnis gelten, wären bei der komplementären Suche nach *laut* als Präposition nicht mindestens 39% *false negatives* zu verzeichnen, so etwa:

***Laut palästinensischer Verfassung*** *endet heute die Amtszeit von Präsident Mahmud Abbas.* [St. Galler Tagbl., 08.01.2009 (S. 5)]

*Verteidigungsminister Barack dagegen sagte* ***laut «Yediot Ahronot»****: «Wir haben unsere Ziele noch nicht erreicht.» Premier Olmert soll angeblich für die Intensivierung der Offensive sein.* [St. Galler Tagbl., 09.01.2009 (S. 5)]

***Laut Ihren Prognosen*** *wird die ALV von 2009 bis 2011 total ein Defizit von fast 5 Milliarden Franken produzieren und Ende 2011 mit fast 9 Milliarden Franken verschuldet sein.* [St. Galler Tagbl., 09.01.2009 (S. 7)]

*Die Zahl der Sturzverletzungen hat sich* ***laut Renato Werndli****, Arzt in Eichberg, sogar verdoppelt.* [St. Galler Tagbl., 09.01.2009 (S. 29)]

Insgesamt also ein wenig befriedigendes Ergebnis.

### Connexor

Mit 9973 korrekten Fundstellen unter 10000 zufällig aus 271608 Treffern ausgewählten Belegen übertrifft Connexor sogar geringfügig das positive Ergebnis des TreeTaggers. Weit größer noch ist der Vorsprung von Connexor allerdings auch bei den *false negatives*: Bei einem zufällig aus 185456 Treffern ausgewählten Sample von 10000 Belegen für die Wortform *laut* nicht als Präposition fanden sich über 8800 Belege, bei denen *laut* eindeutig als Präposition auszuzeichnen gewesen wäre, so etwa in diesen Sätzen:

*Das Prüfungsergebnis lässt sich* ***laut einer Pressemitteilung*** *zeigen.* [St. Galler Tagbl., 22.01.2009 (S. 33)]

*Ziel ist es* ***laut dieser Meldung****, die neue Saison 2009/10 im Sommer mit einem neuen Trainer zu starten.* [St. Galler Tagbl., 25.04.2009 (S. 43)]

*Welche Versicherungen benötigt werden, hängt* ***laut Experten*** *von den Lebensumständen ab.* [Braunschw. Z., 27.08.2009]

*Kennedy-Allee beschweren sich* ***laut SPD*** *über die unhaltbaren Zustände in ihrer Straße.* [Braunschw. Z., 26.08.2009]

Insgesamt betrachtet kann die Erkennungsleistung deshalb nicht befriedigen.

## 4.3 Fazit

Wer erwartet, dass ihm die hier betrachteten Tagger bei der Evaluation von Aussagen zur deutschen Grammatik und zu Variationsphänomenen mit niedriger Frequenz von großer Hilfe sein könnten, wird sich enttäuscht sehen. Man mag zwar vermuten, dass die oft geringe Zahl an *false positives* brauchbare Untersuchungen zur Variation – etwa der Rektion einzelner Präpositionen – zulässt, doch bevor man davon ausgeht, sollte man zumindest prüfen, ob nicht bestimmte Varianten bevorzugt unter den *false negatives* zu finden sind.

| ***eingangs/Eingangs*** | **TreeTagger** (für *eingangs/Eingangs*) | **Connexor** (nur *eingangs*) |
|---|---|---|
| als Präposition | 4010 | 1693 |
| *false positives* | > 72% | > 51% |
| nicht als Präposition | 1454 | 3770 |
| *false negatives* | 0,6% | > 8,3% |

| ***fernab*** | **TreeTagger** | **Connexor** |
|---|---|---|
| als Präposition | 0 | 0 |
| nicht als Präposition | 4104 (alle als Adverbien) | 4069 |
| *false negatives* | >95% | > 87% |

| ***halber*** | **TreeTagger** | **Connexor** |
|---|---|---|
| als Präposition (bzw. Postposition) | 1764 | 1436 |
| *false positives* | < 5% | 3% |
| nicht als Präposition | 4829 | 5138 |
| *false negatives* | 1,3% | 6,9% |

| ***jenseits*** | **TreeTagger** | **Connexor** |
|---|---|---|
| als Präposition | 21193 | 14135 |
| *false positives* | 0,28%* | 0,09%* |
| nicht als Präposition (für *jenseits/Jenseits*) | 7609 | |
| *false negatives* | > 60% | > 60%* |

| *laut* | TreeTagger | Connexor |
|---|---|---|
| als Präposition | 398 500 | 271 608 |
| *false positives* | 0,64%* | 0,27%* |
| nicht als Präposition | | 185 456 |
| *false negatives* | > 39%* | 88%* |

* aus Zufallsauwahl von 10 000

Tab. 6: Präpositionenerkennung mit TreeTagger und Connexor

Wer gar die Entdeckung bislang verkannter Strukturen durch Recherchen in getaggten Textkorpora für möglich hielt, wird erkennen, dass sich zwar Strukturen finden, die in „intelligenten" Grammatiken nicht beschrieben werden, dass es sich dabei jedoch weit bis weitest gehend um Erscheinungen handelt, die fehlerhaften Annotationen geschuldet sind, was sie ohne intensive Überprüfung unbrauchbar macht.

Intensive Überprüfung könnte jedoch das Schlüsselwort für eine dennoch sinnvolle Nutzung getaggter Korpora sein, denn die echten und die vermeintlichen Fehler der Tagger sind keinesfalls reine Zufallsprodukte, sondern haben System. Sie gründen zum einen in einer unzureichenden Auswertung der Kontexte, in denen die zu taggenden Wortformen stehen, zum anderen aber auch in prinzipiellen Schwierigkeiten bei der Bestimmung von Wortklassen – Schwierigkeiten, wie sie sich bei allen Grammatiken zeigen. Typische Beispiele hierfür sind etwa die Unterscheidung von

- adverbialem und prädikativem Adjektiv:
  *Der Schnitt war tief. – Der Schnitt ging tief.*
- Adverb und Fokuspartikel:
  *Das kränkte mich sehr. – Das kränkte mich schon.*
- prädikativem Adjektiv und Präposition:
  *Ich war damals noch ledig. – Ich war endlich aller Sorgen ledig.*

Eine genauere Betrachtung von Erscheinungen, bei denen Tagger zu anderen Entscheidungen kommen als von Theoretikern erwartet, ist nicht allein für Korpuslinguisten von Interesse, sie kann auch für eher traditionalistisch eingestellte Grammatiker als heuristische Methode von Nutzen sein, wenn sie ihre Analysen ernstlich empirisch überprüfen wollen.

# 5. Maschinelles Lernen zur Vorhersage von Fugenelementen in nominalen Komposita

Diese exemplarische Studie zeigt, wie ein Verfahren des maschinellen Lernens eingesetzt werden kann, um Regeln für die Wahl von Fugenelementen in nominalen Komposita aufzudecken. Auf die Basis eines Trainingskorpus von über 400 000 Komposita wird der Algorithmus C4.5 angewandt, um einen sogenannten „Entscheidungsbaum" zu generieren, der die Fugenelemente mit einer hohen Trefferquote vorhersagt. Es wurde versucht, diesen Entscheidungsbaum linguistisch zu deuten, um bestehende Hypothesen über die Wahl von Fugenelementen zu prüfen.

## 5.1 Fragestellung

Die Bildung von Fugenelementen in deutschen nominalen Komposita ist ein Forschungsthema mit langer Tradition (z.B. Ortner et al. 1991; Fuhrhop 1996; Donalies 2011), trotzdem ist es schwierig, konsistente Regeln abzuleiten und die beobachtbare Variation zu erklären. Wichtig ist eine empirische Grundlage, um anhand einer großen Datenmenge Hypothesen über Regularitäten testen zu können. Im Folgenden wollen wir auf der Basis des Deutschen Referenzkorpus des Instituts für Deutsche Sprache (Kupietz et al. 2010) und damit anhand von etwa 400 000 unterschiedlichen Komposita das Thema statistisch über ein Verfahren des maschinellen Lernens, ein KDD-Verfahren (Knowledge Discovery in Databases), angehen: Es ist das Ziel, durch die Berücksichtigung einer Reihe von morphologischen und phonetischen Eigenschaften der Komposita einen Entscheidungsbaum zu modellieren, der Vorhersagen über die Wahl des Fugenelements anhand einer Reihe von hintereinander gelagerten Entscheidungen aufgrund der Eigenschaften der Komposita treffen kann.

Unser Anspruch besteht nicht darin, völlig neue Regeln für das Auftreten von Fugenelementen zu entdecken.[1] Dies wäre aufgrund unserer Herangehensweise auch gar nicht möglich, da für die Modellierung nur solche Eigenschaften berücksichtigt wurden, die auch in existierenden Hypothesen zum Auftreten von Fugenelementen eine Rolle spielen. Vielmehr ist die Methodik, d.h. die Anwendung eines statistischen Lernverfahrens und die Zugrundelegung einer

[1] Ebenso wenig besteht unser Anspruch darin, anhand der mit Hilfe des Entscheidungsbaums formulierbaren Regeln mental repräsentiertes grammatisches Wissen abzubilden. Hier soll es vielmehr um Regeln in einem allgemeinen Sinne gehen, d.h. um linguistische Regeln, anhand derer sich die Distribution von Fugenelementen erklären lässt.

breiten Datenbasis von 407 865 unterschiedlichen Komposita (Lemma-Types), als innovativ zu beurteilen.

Zum einen sollte sich durch die Arbeit mit einem Entscheidungsbaum und durch die ausschließliche Berücksichtigung statistisch signifikanter Daten auf optimale Weise zeigen lassen, welche Kombinationen von Regeln ausschlaggebend für die Gestaltung der Kompositionsfuge sind. Schließlich scheint sich die vermeintliche Fugen-Systematik „nur durch eine Verzahnung der verschiedenen Kriterien" (Fuhrhop 2000: 206) und nicht etwa durch ein Kriterium allein erschließen zu lassen (vgl. Fuhrhop 1996: 525).

Zum anderen kann anhand unseres Entscheidungsbaums überprüft werden, ob die Gestaltung der Kompositionsfuge überhaupt an bestimmte Regeln oder Regelverzahnungen gekoppelt ist,[2] oder ob die „Setzung oder Unterlassung" von Fugenelementen als eine „Frage des Sprachgebrauchs, der Konvention [und] der Üblichkeit" (Fleischer 1971: 117) zu betrachten und somit nicht durch eine Verzahnung von Regeln modellierbar ist. Nicht ohne Grund stellen Wellmann et al. bereits 1974 für den Wortbildungsprozess der Komposition fest, sie sei „ein besonders aufschlußreiches Prüffeld für die Diskussion der Linguisten, auf welche Weise und inwieweit sich sprachliche Vorgänge adäquat in Regeln fassen lassen" (Wellmann et al. 1974: 358). Um die Relevanz von Eigenschaften als Einflussfaktoren für die Fugenbildung einer großen Menge von Komposita berechnen zu können, müssen alle Komposita, die als Datengrundlage dienen, nach diesen Eigenschaften klassifiziert werden. Dies kann nur automatisch geschehen, denn es wäre unmöglich, für über 400 000 Komposita manuell zu bestimmen, welches Suffix, welche Wortart etc. das Erstglied bzw. das Zweitglied aufweisen. Daher können wir auch nur Eigenschaften berücksichtigen, die maschinell bestimmbar sind.

Zur Bestimmung der Eigenschaften verwendeten wir die lexikalische Datenbank CELEX (Baayen et al. 1995), die eine Reihe von morphologischen, orthographischen und phonologischen Merkmale für die darin enthaltenen Lexeme nennt. Eigenschaften, die in dieser Datenbank nicht genannt werden und so nur schwierig maschinell bestimmbar wären, müssen wir ignorieren, dazu gehören z.B. semantische Merkmale.

[2] Ramers kommt in diesem Zusammenhang zu dem Schluss, „daß die Wahl des Fugenelements in Komposita keineswegs willkürlich ist, sondern klaren grammatischen Prinzipien folgt" (Ramers 1997: 44).

Linguistische Erkenntnisse über den Gebrauch von Fugenelementen in Komposita zu gewinnen, ist nur eines der beiden Ziele unseres Ansatzes. Mindestens genauso wichtig ist uns das methodische Experiment: Gelingt es mit maschinellen Lernverfahren, linguistisch sinnvolle Regeln über die Verwendung von Fugenelementen abzuleiten?

## 5.2 Knowledge Discovery in Databases (KDD)

Große Datenmengen wie die IDS-Textkorpora können manuell nicht gewinnbringend analysiert werden. Schon wenn man nur einzelne Belege für ein bestimmtes Phänomen heraussuchen möchte, ist man auf die Hilfe von computergestützten Abfragen angewiesen. Sollen jedoch alle Belege eines Phänomens betrachtet und systematisch ausgewertet werden, so ist man – abgesehen von sehr seltenen Phänomenen mit nur wenigen Belegen – auf Prozesse und Verfahren angewiesen, die mit „Knowledge Discovery in Databases" (KDD), „Data-Mining" oder „Maschinelles Lernen" bezeichnet werden.

Das Ziel von KDD ist, in meist sehr großen Datenbeständen mit mathematisch-statistischen Methoden neue, bisher unbekannte fachliche Zusammenhänge zu erkennen. In der Praxis werden die Begriffe KDD und Data-Mining häufig synonym gebraucht, obwohl Data-Mining streng genommen nur den Modellierungsschritt, also das Lernen von Regularitäten durch den Algorithmus, im KDD-Prozess bezeichnet. Beim maschinellen Lernen werden auf vorhandenen Daten statistische Modelle trainiert, die anschließend zur Vorhersage (z.B. Klassifikation neuer Daten) eingesetzt werden können. Viele der eingesetzten Algorithmen können sowohl für Data-Mining als auch für maschinelles Lernen verwendet werden.

Der KDD-Prozess besteht normalerweise aus folgenden Schritten (vgl. Shearer 2000):

- Aufgabenverständnis: Ziele und grobe Vorgehensweise festlegen,
- Datenverständnis: Zusammenstellung der benötigten Datenquellen und Einarbeitung in deren Besonderheiten,
- Datenaufbereitung: Erstellung der endgültigen Datenmenge für die Modellierung (Auswahl der Fälle und Attribute, gegebenenfalls Berechnung/Umrechnung von Werten),
- Modellierung: Anwendung geeigneter Verfahren zur Datenanalyse und Modellbildung (meist gibt es mehrere passende Verfahren mit mehreren möglichen Parameterkombinationen),

- Evaluation: Auswahl des Modells, das die Aufgabenstellung am besten erfüllt,
- Einsatz: Interpretation des Modells (KDD) oder Anwendung des Modells auf neue Daten (maschinelles Lernen).

Erfahrungsgemäß werden mindestens 80% der Zeit für die ersten drei Schritte aufgewendet.

**Aufgabenverständnis:** Der grundlegende Schritt des KDD-Prozesses ist die Zielfestlegung: Was soll überhaupt modelliert werden und wozu dient die Modellierung? Bei der Modellierung der Fugenelemente geht es darum, den Regeln, die den Gebrauch von Fugenelementen in zweigliedrigen nominalen Komposita des Deutschen steuern, auf die Spur zu kommen. Um dies zu erreichen, muss das im Modell enthaltene Wissen explizit repräsentiert, also für den Menschen „lesbar" sein, damit ein Experte es interpretieren kann. Daher bietet sich hier an, so genannte Entscheidungsbäume zu trainieren.

**Datenverständnis und Datenaufbereitung:** Nach der Festlegung des Ziels wird ermittelt, welche Daten zur Modellierung zur Verfügung stehen. Diese Daten müssen auf ihre Besonderheiten hin untersucht werden, um spätere Überraschungen zu minimieren. Wenn die Basisdaten für die Modellierung nicht ausreichen, müssen Daten aus weiteren Quellen integriert werden. So wurden beispielsweise zusätzlich zu den aus den Korpora gewonnenen Daten für die Fugenelementanalyse noch Daten aus der lexikalischen Datenbank CELEX (Baayen et al. 1995) extrahiert und hinzugefügt.

**Modellierung:** Bei der Modellierung berechnet ein Algorithmus aus einer Menge von Trainingsdaten ein Modell, das dann auf neue Daten angewendet werden kann, um bestimmte Werte vorherzusagen. Die Algorithmen unterscheiden sich unter anderem bezüglich der Art und Anzahl der vorhergesagten Werte, der Art der Trainingsdaten, überwachter vs. unüberwachter Methoden und der Interpretierbarkeit des Modells.

Bei überwachten Methoden („supervised learning") enthalten die Trainingsdaten auch die tatsächliche Kategorie, die man für neue Daten vorhersagen können möchte. Bei der Modellierung der Fugenelemente kennen wir beispielsweise für alle aus den IDS-Korpora extrahierten zweigliedrigen Komposita das jeweilige Fugenelement, daher können wir auf diese Daten überwachte Methoden anwenden. Bei unüberwachten Methoden („unsupervised learning"), wie z.B. Clustering, werden die Trainingsdaten vom Lernalgorithmus

in (eine meist vorgegebene Anzahl von) Kategorien eingeteilt. Diesen wird meist im Nachgang manuell eine passende Bezeichnung zugeordnet.

Wie bereits erwähnt sollte das Modell für den vorliegenden Zweck lesbar und interpretierbar sein. Beispielsweise erlauben künstliche neuronale Netze keinen Einblick in ihre durch die Modellierung erlernte Struktur und können daher nicht vom Menschen interpretiert werden. Entscheidungsbäume hingegen sind – vorausgesetzt sie sind nicht zu groß – gut les- und interpretierbar.

**Evaluation:** Die trainierten Modelle sollten nicht anhand der Trainingsdaten, sondern anhand separater Testdaten, die nicht zum Training verwendet wurden, evaluiert werden. Eine bewährte Methode ist die k-fache Kreuzvalidierung („k-fold cross validation"). Dazu werden die Trainingsdaten in $k$ Teilmengen aufgeteilt. In $k$ Trainings- und Testdurchläufen wird jeweils eine der $k$ Teilmengen als Testdatensatz verwendet, und die verbleibenden $k\text{-}1$ Teilmengen bilden den Trainingsdatensatz. Der Durchschnitt aus den Evaluationsmaßen der $k$ Einzeldurchläufe bildet das Gesamtmaß.

Bei nominalen Vorhersagewerten, wie sie in der vorliegenden Analyse mit den unterschiedlichen Fugenelementen vorliegen, kommen meist mehrere Evaluationsmaße zum Einsatz. Dafür wird für jede vorherzusagende Klasse X in den Testdaten Folgendes gezählt:

- richtig positiv ($r_p$): Anzahl der Fälle, bei denen das Modell vorhersagt, dass der Fall zur Klasse X gehört, und dies auch der Realität entspricht;
- richtig negativ ($r_n$): Anzahl der Fälle, bei denen das Modell vorhersagt, dass der Fall nicht zur Klasse X gehört, und dies auch der Realität entspricht;
- falsch positiv ($f_p$): Anzahl der Fälle, bei denen das Modell vorhersagt, dass der Fall zur Klasse X gehört, aber dies nicht der Realität entspricht;
- falsch negativ ($f_n$): Anzahl der Fälle, bei denen das Modell vorhersagt, dass der Fall nicht zur Klasse X gehört, aber dies nicht der Realität entspricht.

In einer Konfusionsmatrix angeordnet sieht es folgendermaßen aus:

| | **Vorhersage: Klasse X** | **Vorhersage: nicht Klasse X** |
|---|---|---|
| **Realität: Klasse X** | $r_p$ | $f_n$ |
| **Realität: nicht Klasse X** | $f_p$ | $r_n$ |

Tab. 1: Konfusionsmatrix

Mit diesen Werten lassen sich nun folgende Evaluationsmaße berechnen:

- Genauigkeit („precision") = $r_p : (r_p + f_p)$
- Trefferquote („recall") = $r_p : (r_p + f_n)$
- F-Maß = 2 · Genauigkeit · Trefferquote : (Genauigkeit + Trefferquote)

Alle drei Maße können Werte zwischen 0 und 1 annehmen – je größer, desto besser das Modell. Dabei sind Genauigkeit und Trefferquote oft voneinander abhängig: Die Trefferquote kann hoch, die Genauigkeit jedoch niedrig sein, wenn zwar alle Fälle der Klasse X gefunden werden, dabei jedoch auch viele Fälle der Klasse „nicht X" fälschlicherweise dabei sind. Umgekehrt kann die Genauigkeit hoch, die Trefferquote jedoch niedrig sein, wenn zwar alle der Klasse X zugeordneten Fälle tatsächlich der Klasse X angehören, dabei aber viele Fälle nicht dabei sind, die eigentlich auch zu Klasse X gehören würden. So möchte man das bestmögliche Verhältnis von Genauigkeit und Trefferquote erreichen. Das F-Maß drückt dieses Verhältnis aus, indem es Genauigkeit und Trefferquote mithilfe des harmonischen Mittels (gegebenenfalls gewichtet) kombiniert.

Ist der Einsatzzweck die Interpretation des Modells durch einen menschlichen Experten, so ist nicht unbedingt das Modell mit den besten Evaluationsmaßen am besten dafür geeignet. Z.B. eignet sich ein kleinerer Entscheidungsbaum mit wenigen Knoten besser zur Interpretation, auch wenn er kein so gutes F-Maß aufweist wie ein Entscheidungsbaum mit vielen Knoten.

**Einsatz:** Je nach Ziel des KDD-Prozesses besteht der Einsatz des Modells in seiner Interpretation durch einen menschlichen Experten und dem damit verbundenen Erkenntnisgewinn oder in der Anwendung auf neue Daten, um diese zu klassifizieren bzw. andere Vorhersagen zu treffen.

## 5.3 Datenaufbereitung

### 5.3.1 Datenextraktion

Grundlage für die Modellierung der Fugenvorhersage ist ein Auszug von zweigliedrigen nominalen Komposita aus dem DeReKo (Kupietz et al. 2010).[3] Die Daten wurden mittels einer automatischen Analyse der morphosyntaktisch

[3] Dieser Auszug von Komposita wurde bereits für eine im Rahmen des Projekts „Korpusgrammatik" entstandene, aber anders gelagerte Analyse von Zusammensetzungen von Elke Donalies (2011) erstellt, wo allerdings nicht nur zweigliedrige Komposita berücksichtigt wurden. Darüber hinaus ist die Analyse von Donalies weniger daten-, sondern stärker hypothesengeleitet.

annotierten Version des DeReKo erhoben. Das DeReKo steht in verschiedenen morphosyntaktisch annotierten Versionen zur Verfügung.[4] Zum Aufbau der Schnittstellen-Datenbank wurde die mit dem „Machinese"-Tagger der Firma „Connexor"[5] annotierte Version verwendet, denn der Tagger nimmt bei Zusammensetzungen eine morphologische Analyse vor, wie im folgenden Ausschnitt der Annotation von *Altersgruppe* zu sehen ist:

```
<token pos="3671970" len="12">
<text>Altersgruppe</text>
<lemma>alter gruppe</lemma>
<tags syntax="@NH" morpho="N"/>
</token>
```

Das Token *Altersgruppe* wird also im Tag `<lemma>` in die Glieder *alter* und *gruppe* aufgetrennt. Die Schnittstelle *s* wird dabei ignoriert.

Mit einem speziell für diesen Zweck programmierten Perl-Skript[6] werden die Connexor-getaggten Korpusdaten ausgelesen. Berücksichtigt werden alle als Nomen klassifizierten Token, die gemäß Lemma-Angabe in der Annotation aus mehreren Gliedern bestehen. Die aufgeführten Glieder (im Beispiel oben: *alter* und *gruppe*) werden nun mit dem tatsächlichen Token verglichen (*Altersgruppe*). Es wird berechnet, welche Operationen notwendig sind, um zum Token zu gelangen[7]. Dabei werden die folgenden Fälle in der genannten Reihenfolge unterschieden:

1) Die Glieder lassen sich über eine der vordefinierten Schnittstellen *s*, *en*, *e*, *n*, *er*, *es*, *a* verbinden (*Pferd* + *e* + *wagen*).
2) Die Glieder lassen sich ohne jegliches weitere Element miteinander verbinden (*Greif* + *vogel*).
3) Eine beliebige andere Zeichenkette ist notwendig, um die Glieder zu verbinden (*Herz* + *ens* + *güte*).

[4] Vgl. http://www.ids-mannheim.de/kl/projekte/korpora/annotationen.html (Stand: 2.8.2013).

[5] Vgl. Kapitel 2.1 und http://www.connexor.eu/technology/machinese/demo/ (Stand: 2.8.2013).

[6] Perl ist eine Programmiersprache, die sich besonders für die Verarbeitung von Textdokumenten eignet (Wall et al. 2000).

[7] Aufgrund dieser „maschinellen Gegebenheiten" wird ein Konzept von Fugenelementen vertreten, nach dem alle Zeichenketten, die sich zwischen zwei Gliedern (Stämmen) eines Kompositums befinden, als Fugenelement betrachtet werden. Für detailliertere Informationen zu den beiden grundsätzlichen Auffassungen von Fugenelementen vgl. Donalies (2011: 7-12).

4) Eine Verkürzung des ersten Glieds in Verbindung mit einem der vordefinierten Schnittstellen ist notwendig (*Hilf + s + konstruktion* [*Hilfe → Hilf + s*]).
5) Verkürzung des ersten Glieds ohne weitere Veränderungen (*Grenz + dienst* [*Grenze → Grenz + dienst*]).
6) Wenn keiner der genannten Fälle zutrifft, wird die Zusammensetzung als unanalysierbar markiert, wobei eine Reihe von häufigen Spezialfällen berücksichtigt werden, die auf orthografische Unterschiede zwischen Lemma und Token wie *-graph-* vs. *-graf-*, *-photo-* vs. *-foto-* etc. zurückgehen.

Generell ignoriert werden Pluralumlaute wie in *Ärztekongress* (*arzt + kongress*): Vor der Analyse werden Umlaute generell ersetzt und die Zusammensetzung *Ärztekongress* würde dann als Fall *Arzt + e + Kongress* klassifiziert.

Neben der Klassifizierung der Zusammensetzungen wird zu jedem gefundenen Token auch die genaue Fundstelle (Korpus und Position im Korpus) abgelegt.

Die extrahierten Komposita werden zu einer aggregierten Liste der unterschiedlichen Komposita (Types) zusammengeführt: Sie enthält alle unterschiedlichen Kombinationen von Erstglied und Zweitglied mit Häufigkeiten und der Fuge oder dem Wert „variabel", wenn die Fuge variiert.

Wir wollen uns also im Folgenden auf Komposita konzentrieren, deren Verfugungsverhalten in den zugrundeliegenden Korpusdaten stabil ist. Zwar werden Komposita, die mit unterschiedlichen Fugen vorkommen, mit dem Wert „variabel" und der Angabe, welche Varianten angetroffen wurden, erfasst, doch beziehen wir die Frequenzverhältnisse dieser Varianten nicht mit in die Analyse ein. Es ist geplant, die bezüglich Fuge variierenden Komposita in einer Folgestudie separat in den Blick zu nehmen.[8]

Nach der oben beschriebenen Extraktion und Aufbereitung der Komposita schränkten wir die Datengrundlage weiter ein auf Komposita, die die folgenden Bedingungen erfüllen:

1) nominale Komposita (Kompositum muss vom Tagger als Nomen klassifiziert sein, nicht jedoch die Glieder),
2) zweigliedrige Komposita,

[8] Vgl. zum Thema der Variation bei der Wahl von Fugenelementen die aktuellen Studien von Nübling/Szczepaniak (2011) und Donalies (2011).

3) keine Bindestrich-Komposita,

4) Frequenz des Kompositums im Korpus ist mindestens fünf

5) Die Glieder des Kompositums müssen in der CELEX-Datenbank vorhanden sein. Vgl. dazu die Ausführungen zur CELEX-Datenbank weiter unten (Kapitel 5.4.1).

Aus dieser Vorgehensweise resultiert eine Liste von 407 865 unterschiedlichen Komposita, die die Datengrundlage für die weiteren Analysen bilden.

### 5.3.2 Fehlerquellen

Es gibt eine ganze Reihe von Fehlerquellen, die zu Fehlanalysen der Komposita führen können. Grundsätzlich können bereits Fehler bei der Tokenisierung (Erkennung der Wortgrenzen) und anderer Verfahren der Korpusaufbereitung einen Einfluss haben. Auf solche Probleme soll an dieser Stelle nicht eingegangen werden. Entscheidender sind Fehlanalysen des Connexor-Taggers: Wie jeder Wortarten-Tagger weist auch der Connexor-Tagger Fehler auf (vgl. Kapitel 4). Fehler im Bereich der morphologischen Analyse wirken sich auf die automatische Bestimmung der Komposita aus. Da es sich beim Connexor-Tagger um ein kommerzielles Produkt handelt, ist die Funktionsweise intransparent und Korrekturen schwierig. Folgende beiden Fehlertypen wurden beobachtet:

1) Kompositum wird nicht als Kompositum erkannt: Das Kompositum *Embryonenforschung* wird vom Tagger nicht als Kompositum erkannt und deshalb nicht in die Glieder *Embryo* und *Forschung* zerlegt. Interessanterweise wird jedoch das Kompositum *Embryoforschung* korrekt zerlegt. Es ist momentan nicht abschätzbar, wie viele ähnliche Fälle von (teilweise) falsch analysierten Komposita es gibt. Es ist jedoch anzunehmen, dass allgemein eher seltene Lexeme falsch analysiert sind.

2) Komposita werden nicht konsistent zerlegt: Das Kompositum *Willensstärke* wird mit dem Erstglied *Willen* analysiert. Die Variante *Willenstärke* hingegen mit *Wille* als Erstglied. Im Falle von *Wille-* führen die unterschiedlichen Analysen zu Verzerrungen bei der Verteilung der Fugen zwischen *ns-* und *s*-Fugen.

Wahrscheinlich hängen diese beiden Fehlertypen zusammen, allerdings kann darüber nur spekuliert werden, da die morphologische Analyse des Taggers nicht ausreichend dokumentiert ist, um die Funktionsweise nachvollziehen zu können.

### 5.3.3 Einflussfaktoren

In der Literatur werden eine Reihe von Hypothesen über die Wahl des Fugenelements aus den Bereichen Morphologie, Phonologie und Semantik genannt (vgl. z.B. Fleischer/Barz 1995; Fuhrhop 1996; Ortner et al. 1991; Duden 2005). Um die große Menge an Komposita bezüglich dieser Einflussfaktoren klassifizieren zu können, müssen die Faktoren maschinell bestimmt werden. Unter Zuhilfenahme der CELEX-Datenbank (Baayen et al. 1995) konnten wir einige potenzielle Einflussfaktoren aus den Bereichen Morphologie und Phonologie berücksichtigen. Unberücksichtigt bleiben semantische Faktoren. Im Detail wurden folgende Faktoren (jeweils von Erst- und Zweitglied) ausgewählt:

1) Morphologie
   - Wortart
   - Flexionsparadigma
   - morphologischer Status
   - falls vorhanden: Suffix (Orthografie und Typus)
   - falls vorhanden: Präfix (Orthografie und Typus)
   - abtrennbares Präfix (ja/nein)

2) Phonologie
   - für den letzten Laut des Erstglieds bzw. den ersten Laut des Zweitglieds:
     - phonetische Umschrift
     - Vokal/Konsonant
     - nur bei Konsonanten:
       - stimmhaft/stimmlos
       - Artikulationsart: Plosiv, Nasal, Frikativ, Affrikate, Liquid, Approximant
       - Artikulationsort: labial, alveolar, palato-alveolar, palatal, velar, uvular, glottal
   - Anzahl der Silben
   - Betonung:
     - erste Silbe (betont/unbetont)
     - vorletzte Silbe (betont/unbetont)

- letzte Silbe (betont/unbetont)
- Position der betonten Silbe (Zahl)

Im Folgenden wird beschrieben, wie die Komposita mit diesen Eigenschaften angereichert wurden.

## 5.4 Anreicherung der Daten

Für die Erstellung der Trainings- und Testdaten muss jeder Fall mit seinen Attributen in eine Vektordarstellung gebracht werden. Bei der Fugenelementanalyse ist jedes Kompositum aus dem Korpus ein Fall. Attribute sind die bekannten Eigenschaften der Fälle, z.B. Silbenzahl des Erstglieds, Wortart des Erstglieds oder der erste Laut des Zweitglieds in phonetischer Umschrift. Bei überwachten Lernverfahren ist eins der Attribute das so genannte Zielattribut, das das zu trainierende Modell vorhersagen soll, hier also das Fugenelement. Zwei Beispielvektoren werden in Tabelle 2 ausschnittsweise dargestellt (die Attribute 'Vektornummer' und 'Kompositum' werden im Training nicht berücksichtigt):

| Vektor-nummer | Kompositum | Silbenzahl des Erstglieds | letzter Laut des Erstglieds | Wortart des Erstglieds | erster Laut des Zweitglieds | ... | Fugen-element |
|---|---|---|---|---|---|---|---|
| 1 | *Evakuierungstest* | 5 | N | N | T | ... | s |
| 2 | *Umweltschutz* | 2 | T | N | S | ... | 0 |

Tab. 2: Ausschnitt zweier Beispielvektoren für die Fugenelementanalyse

Theoretisch kann man alle Attribute, die einem zur Verfügung stehen, in den Vektor aufnehmen und es dem maschinellen Lernalgorithmus überlassen herauszufinden, welche davon die entscheidenden sind. John/Dept (1997) haben jedoch gezeigt, dass das Ergebnis eines maschinellen Lernalgorithmus umso schlechter ist, je mehr irrelevante Attribute die Trainingsdaten enthalten. Besonders problematisch ist hierbei Multikolinearität, d.h. wenn mehrere Attribute stark miteinander korrelieren. Um diese Probleme zu minimieren, sollte man nur Attribute aufnehmen, von denen man annimmt, dass sie einen Einfluss auf das Zielattribut haben. Durch die Berechnung von Korrelationskoeffizienten kann man außerdem ermitteln, welche Attribute stark miteinander korrelieren, und von diesen nur eines im Vektor belassen.

Bei sehr großen Datenmengen kann es auch nötig sein, nicht alle Fälle in die Menge der Trainingsdaten aufzunehmen, da sonst die Modellierung zu viel Zeit in Anspruch nehmen würde. In diesem Fall wird eine stratifizierte Stichprobe gezogen, bei der der relative Anteil der verschiedenen Zielattributsklassen in der Ausgangsdatenmenge und der Stichprobe gleich ist. Genausogut kann es aber nötig sein, die Stichprobe so zu ziehen, dass jede Zielattributsklasse gleich häufig ist. Das kann dann der Fall sein, wenn eine Zielattributsklasse (z.B. die Null-Fuge) im Vergleich zu den anderen Klassen sehr häufig vorkommt und das Modell dazu tendiert, nur diese häufigste Klasse vorherzusagen.

### 5.4.1 CELEX

CELEX (Baayen et al. 1995) ist eine Datenbank mit lexikalischen Informationen zum Niederländischen, Englischen und Deutschen, die in einem Gemeinschaftsprojekt der Universität Nijmegen, des Instituts für Niederländische Lexikologie in Leiden, des Max-Planck-Instituts für Psycholinguistik in Nijmegen und des Instituts für Perzeptionsforschung in Eindhoven entwickelt wurde.

Der deutsche Teil von CELEX enthält 51 728 Grundformen (bei Verben ist dies die Infinitivform und bei Nomen die Nominativ-Singular-Form) und 365 530 flektierte Wortformen, die aus mehreren deutschsprachigen Korpora extrahiert wurden:

- 5,4 Millionen Token aus geschriebenen Texten aus Zeitungen, Belletristik und Sachbüchern („Mannheimer Korpus I und II“ und „Bonner Zeitungskorpus 1“),
- 600 000 Token aus transkribierten Gesprächen („Freiburger Korpus“).

Die Texte und Gespräche wurden zwischen 1949 und 1975 publiziert bzw. aufgenommen und sind am IDS über COSMAS recherchierbar.

### 5.4.2 Beschreibung der Attribute, die aus CELEX stammen

Zu jeder Grundform und jeder flektierten Wortform enthält CELEX verschiedene phonologische, morphologische und syntaktische Informationen verteilt auf mehrere Tabellen. Da die in CELEX enthaltenen Informationen zum Teil automatisch erstellt und nicht komplett manuell überprüft wurden, sind sie in einigen Fällen fehlerhaft. Aus den CELEX Grundform-Tabellen haben wir mit einem für diesen Zweck erstellten Perl-Skript alle oben in Kapitel 5.3.3 aufge-

führten Informationen für alle Grundformen extrahiert und in einer Tabelle zusammengefasst. Im Folgenden beschränken wir uns auf die genaue Darstellung der Attribute, die im beschriebenen Entscheidungsbaum (siehe Kapitel 5.7) eine Rolle spielen.

**Orthografie**: Da der Connexor-Tagger die Kompositumsglieder komplett in Kleinbuchstaben ausgibt, haben wir auch die CELEX-Grundformen in Kleinschreibung transformiert. Außerdem gibt der Connexor-Tagger nur Stammformen aus (also *schwimm* und nicht *schwimmen* wie in CELEX). Daher haben wir die Verbgrundformen automatisch in Stammformen umgewandelt. Aus diesen beiden Gründen gibt es in der neu erstellten Tabelle ambige Grundformen. Z.B. stammen die Informationen zur Grundform mit der Orthografie *bild* zum einen vom Nomen *Bild*, zum anderen vom Verb *bilden*. Daher hat diese Grundform zwei durch Semikolon getrennte Werte für die Wortart, nämlich „N; V".

**Wortart**:[9] Wie in Tabelle 3 zu sehen, unterscheidet CELEX im Deutschen zehn verschiedene Wortarten.

| Kodierung | Wortart | Beispiel | Häufigkeit |
|---|---|---|---|
| A | Adjektiv | *klein* | 9 855 = 19,1% |
| B | Adverb | *anstandshalber* | 1 284 = 2,5% |
| C | Konjunktion | *und* | 78 = 0,2% |
| D | Artikel | *das* | 2 = 0,004% |
| I | Interjektion | *ach* | 37 = 0,1% |
| N | Nomen | *Haus* | 30 715 = 59,4% |
| O | Pronomen | *ich* | 116 = 0,2% |
| P | Präposition | *von* | 108 = 0,2% |
| Q | Quantor/Numeral | *mehr, sechs* | 133 = 0,3% |
| V | Verb | *abstellen* | 9 400 = 18,2% |

**Tab. 3: Kodierung der in CELEX unterschiedenen Wortarten. Die Häufigkeit gibt die Anzahl der Einträge in der CELEX-Grundformen-Datei wieder.**

[9] Die Wortart spielt im weiter unten präsentierten Entscheidungsbaum zwar keine unmittelbare Rolle, die Codes werden aber im Flexionsparadigma ebenfalls verwendet.

**Flexionsparadigma:** CELEX unterscheidet die in Tabelle 4 dargestellten Flexionsparadigmen. Wie in Tabelle 5 und Tabelle 6 zu sehen, werden die nominalen Flexionsparadigmen besonders differenziert und es werden im Singular 7 und im Plural 13 Flexionsklassen unterschieden. Diese Klassifikation ist demnach detaillierter als andere übliche Flexionsparadigmen wie z.B. in grammis (grammis, „Nomen").

| Flexionsparadigma | Bedeutung | Beispiel | Häufigkeit |
|---|---|---|---|
| A | adjektivische Flexion für Nomen | *Angestellte* | 192 |
| I | flektiert, aber kein Paradigma verfügbar | *abermalig* | 9861 |
| U | unflektiert | *aber* | 1752 |
| i | irreguläres Verb | *abbeißen* | 2039 |
| r1 | reguläres Verb | *abbuchen* | 4369 |
| r2 | reguläres Verb auf -d, -t oder -(Plosiv/Frikativ)+(m/n) | *abzeichnen* | 846 |
| r3 | reguläres Verb auf -@r | *abmagern* | 684 |
| r4 | reguläres Verb auf -@l | *abschütteln* | 664 |
| r5 | reguläres Verb auf -(Vokal) | *anflehen* | 222 |
| r6 | reguläres Verb auf -(Sibilant) | *abhetzen* | 576 |
| S[0-6] | nominales Singularflexionsparadigma | | 30526 |
| P[0-10][U] | nominales Pluralflexionsparadigma | | 30526 |

Tab. 4: Kodierung der in CELEX unterschiedenen Flexionsparadigmen

| Code | Maskulina | Feminina | Neutrum |
|---|---|---|---|
| S0 | Pluralia Tantum | | |
| S1 | *der Wald*; *-(e)s* | | *das Brot*; *-(e)s* |
| S2 | *der Bär*; *-(e)n* | | |
| S3 | | *die Bar*; - | |
| S4 | *der Bus*; *-ses* | | *das Zeugnis*; *-ses* |
| S5 | *der Buchstabe*; *-ns* | | |
| S6 | | | *das Herz*; *-ens* |

Tab. 5: Kodierung nominaler Singularflexionsparadigmen in CELEX

| Code | Pluralform | Code | Pluralform |
|---|---|---|---|
| P0 | Singularia Tantum | P4U | *die Dächer*; *-n* |
| P1 | *die Stoffe*; *-n* | P5 | *die Autos*; - |
| P1U | *die Bäume*; *-n* | P6 | *die Freundinnen*; - |
| P2 | *die Esel*; *-n* | P7 | *die Geheimnisse*; *-n* |
| P2U | *die Äpfel*; *-n* | P8 | *die Maxima*; - |
| P3 | *die Bauern*; - | P9 | *die Gymnasien*; - |
| P4 | *die Felder*; *-n* | P10 | andere Wörter |

Tab. 6: Kodierung nominaler Pluralflexionsparadigmen in CELEX

**Suffixe und Präfixe:** In der Datenbank sind zu jedem Lexem Informationen über Suffixe und Präfixe vorhanden. In unserem Entscheidungsbaum spielen nur Suffixe eine bedeutende Rolle, etwa Suffixe wie *-ung*, *-schaft*, *-in* etc. Allerdings muss angemerkt werden, dass die in CELEX verzeichneten Suffixe nicht immer auch im morphologisch engen Sinn Suffixe sind, was bei der Interpretation des Entscheidungsbaums noch deutlich werden wird.

**Letzter Laut und letzte Silbe:** Die Datenbank verzeichnet einerseits den Typ des letzten Lautes (Vokal oder Konsonant) als auch den Laut und die Silbe. Zudem ist angegeben, ob die letzte Silbe betont ist. In CELEX sind diese Informationen in der DISC-Schreibweise (Burnage 1990) wiedergegeben, die wir aber in der Darstellung unseres Entscheidungsbaums in IPA-Schreibweise wiedergegeben haben.

## 5.5 Training des Entscheidungsbaums zur Vorhersage von Fugenelementen

Ein Entscheidungsbaum ist ein gerichteter azyklischer Graph, mit dem Daten automatisch klassifiziert werden können. Er besteht immer aus einem so genannten Wurzelknoten an der Spitze des Entscheidungsbaums, beliebig vielen inneren Knoten (es gibt auch Entscheidungsbäume ohne innere Knoten) und mindestens zwei Blättern. Um einen Fall zu klassifizieren, geht man vom Wurzelknoten abwärts über innere Knoten bis man ein Blatt erreicht. An jeden Knoten wird ein Attribut abgefragt (z.B. „Endet das Erstglied mit einem Konsonanten?"). Je nach Antwort folgt man einem unterschiedlichen Zweig. Das Blatt enthält schließlich die Klassifikation (z.B. „Fugenelement ist *s*"). Vgl. zur Illustration Abbildung 1.

Entscheidungsbäume können mit maschinellen Lernverfahren auf der Basis von Trainingsdaten automatisch erstellt werden. Wir haben im Folgenden den von Quinlan (1993) entwickelten Algorithmus C4.5 verwendet, der in der Software-Suite WEKA (Witten/Frank 2005) als J48 in Java implementiert ist. Bei der Konstruktion des Entscheidungsbaums geht der Algorithmus folgendermaßen vor: Ziel ist, in den Blättern des Entscheidungsbaums nur noch Fälle einer Klasse (z.B. „Fugenelement ist *s*") vorzufinden. Dazu werden zunächst alle Fälle des Trainingsdatensatzes betrachtet. Jedes Attribut wird daraufhin getestet, ob es die Datenmenge in möglichst „reine" Gruppen aufteilt was das Zielattribut betrifft. Das Maß, das der C4.5-Algorithmus zur Bewertung der Attribute verwendet, heißt Kullback-Leibler-Divergenz (Kullback/Leibler 1951). Das Attribut mit dem höchsten Wert wird ausgewählt und der Trainingsdatensatz in Teilmengen nach diesem Attribut aufgeteilt. Für jede Teilmenge werden wiederum die übrigen Attribute anhand der Kullback-Leibler-Divergenz bewertet und nach der Auswahl des Attributs mit dem höchsten Wert in weitere Teilmengen aufgeteilt. Dieser Prozess wiederholt sich, bis eine Teilmenge nur noch Fälle einer Klasse enthält oder die vom Benutzer vorgegebene minimale Anzahl von Fällen pro Blatt erreicht ist. Das Blatt erhält dann die Klasse mit den meisten Fällen. Da die Gefahr besteht, dass der so trainierte Entscheidungsbaum zwar die Trainingsdaten optimal klassifiziert, unbekannte Daten jedoch nicht gut klassifiziert (Gefahr der Überanpassung an die Trainingsdaten), wird der Entscheidungsbaum in einem letzten Schritt noch „zurückgeschnitten" („pruning").

Für die Fugenelementanalyse haben wir mehrere Bäume trainiert, im Folgenden werden jedoch nur die Resultate zu einem der Bäume dargestellt. Die Trainingsdaten dieses Baums enthalten jede Erstglied-Zweitglied-Folge nur ein einziges Mal, egal wie häufig sie in den Korpora vorkommt (vgl. dazu auch Kapitel 5.3.1). Wenn eine Erstglied-Zweitglied-Folge in den Korpora mit verschiedenen Fugenelementen vorkommt (so gibt es z.B. sowohl *Abfahrtszeit* als auch *Abfahrtzeit*), so wird als Zielattribut ein variables Fugenelement angegeben (in diesem Fall „[var_s_0]").

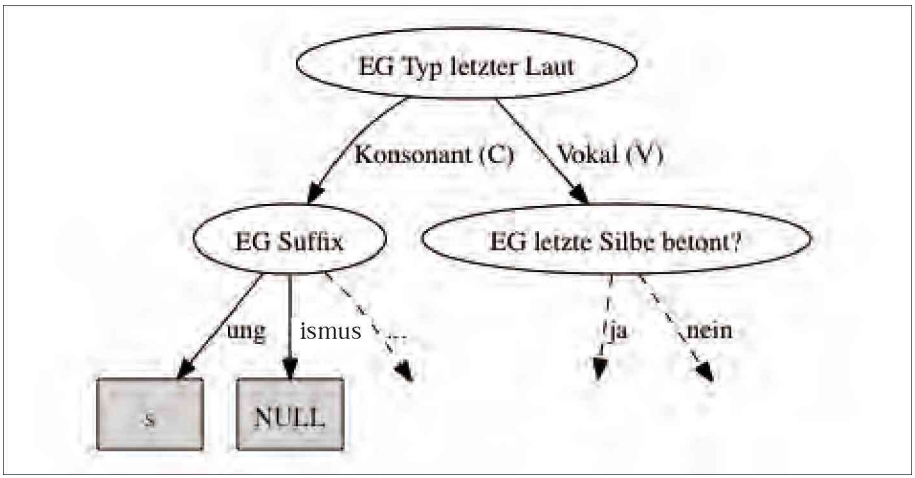



Abb. 1: Ausschnitt aus einem Entscheidungsbaum zur Vorhersage von Fugenelementen (EG = Erstglied)

## 5.6 Evaluation

Im Folgenden werden die Evaluationsmaße für den Baum dargestellt, die bei zehnfacher Kreuzvalidierung mit einem Stoppkriterium von mindestens zwei Fällen pro Blatt erreicht wurden. Der Baum, der im Interpretationsabschnitt präsentiert wird, ist aus Gründen der Interpretierbarkeit mit einem groben Stoppkriterium (mindestens 500 richtige oder falsche Fälle pro Blatt) trainiert und aus Gründen der Übersichtlichkeit vereinfacht.

In Tabelle 7 werden die Evaluationsmaße Genauigkeit, Trefferquote und F-Maß für den Entscheidungsbaum, der auf allen Erstglied-Zweitglied-Folgen trainiert wurde, dargestellt. Wenn die Erstglied-Zweitglied-Folge ein variables Fugenelement hat, so wurden alle Fälle als korrekte Vorhersage gewertet, bei denen mindestens eins der möglichen Fugenelemente vorhergesagt wurde. Wenn also das Fugenelement *s* oder 0 (=Null-Fuge) sein kann, so wurden die drei Voraussagen *s*, 0 und „[var_0_s]" als korrekt gewertet.

Insgesamt werden durchschnittlich 95% der Testdaten korrekt klassifiziert, wie Tabelle 7 zeigt (F-Maß). Die häufigsten Fugen, die Null- und die *s*-Fuge, werden in 97% bzw. 96,5% der Fälle korrekt vorausgesagt, die ebenfalls häufige *n*- und *en*-Fuge allerdings zu 89 bzw. 87% der Fälle. Besonders schlecht wird die Tilgung von *e* (z.B. in *Mietwohnung; Miete + Wohnung*) vorausgesagt (65%).

Tabelle 7 zeigt darüber hinaus auch die Verteilung der Komposita auf die unterschiedlichen Fugenelemente. Obwohl die Verteilung tendenziell anderen Analysen ähnelt (vgl. für eine Übersicht Donalies 2011: 32), gibt es Differenzen, die mit den unterschiedlichen Korpuszusammensetzungen und Berechnungs- bzw. Erhebungsmethoden zusammenhängen. Insbesondere die Unterschiede zur Verteilung, die in Donalies (ebd.) dargestellt ist und die eigentlich auf die gleiche Datenbasis zurückgeht, können auf die differierenden Auswahlkriterien der vorliegenden Studie zurückgeführt werden. So berücksichtigten wir nur zweigliedrige Komposita, deren Glieder in CELEX vorhanden sind (vgl. Kapitel 5.3). Gleichwohl zeigt sich auch in unseren Daten, dass die Null-Fuge den größten Anteil ausmacht, gefolgt von der *s*- und der *(e)n*-Fuge.

| Fugenelement | Anzahl Fälle | Anteil in % | Genauigkeit | Trefferquote | F-Maß |
|---|---|---|---|---|---|
| 0 | 247 217 | 60,61% | 0,967 | 0,976 | 0,971 |
| *s* | 87 186 | 21,38% | 0,953 | 0,979 | 0,965 |
| *n* | 37 933 | 9,30% | 0,843 | 0,933 | 0,886 |
| *en* | 10 752 | 2,64% | 0,910 | 0,830 | 0,868 |
| variables Fugenelement | 7 985 | 1,96% | 0,952 | 0,979 | 0,966 |
| *-e* (=Tilgung des *e*) | 6 874 | 1,69% | 0,725 | 0,582 | 0,646 |
| *es* | 3 076 | 0,75% | 0,648 | 0,922 | 0,761 |
| *er* | 2 906 | 0,71% | 0,829 | 0,712 | 0,766 |
| *e* | 2 504 | 0,61% | 0,729 | 0,749 | 0,739 |
| *um → en* | 1 113 | 0,27% | 0,975 | 0,980 | 0,978 |
| *nen* | 167 | 0,04% | 0,906 | 0,982 | 0,943 |
| *ns* | 152 | 0,04% | 0,849 | 1,000 | 0,918 |
| **alle** | **407 865** | **100,00%** | **0,941** | **0,958** | 0,950 |

**Tab. 7: Evaluationsmaße für den ersten Entscheidungsbaum aufgeteilt nach Fugenelement**

Doch nun zurück zu den Daten der Evaluation: Einen besseren Überblick über die problematischen Voraussagen bieten Konfusionsmatrizen. Sie stellen dar, welches Fugenelement vorausgesagt wurde und welches Fugenelement das Kompositum tatsächlich aufwies.

| | vorhergesagtes Fugenelement | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| tatsächliches Fugenelement | 0 | *-e* | *e* | *en* | *um → en* | *er* | *es* | *n* | *nen* | *ns* | *s* | Summe |
| **0** | 241 273 | 234 | 317 | 544 | 10 | 105 | 531 | 2924 | 15 | 0 | 1083 | 247036 |
| *-e* | 148 | 3999 | 0 | 0 | 0 | 0 | 0 | 2723 | 0 | 0 | 0 | 6870 |
| *e* | 486 | 0 | 1875 | 15 | 0 | 5 | 54 | 0 | 0 | 0 | 61 | 2496 |
| *en* | 1250 | 0 | 7 | 8928 | 0 | 0 | 2 | 1 | 0 | 0 | 545 | 10733 |
| *um → en* | 0 | 0 | 0 | 0 | 1091 | 0 | 0 | 0 | 0 | 0 | 21 | 1112 |
| *er* | 522 | 0 | 58 | 1 | 0 | 2070 | 148 | 0 | 0 | 0 | 96 | 2895 |
| *es* | 180 | 0 | 0 | 0 | 0 | 43 | 2835 | 0 | 0 | 0 | 6 | 3064 |
| *n* | 1385 | 1106 | 0 | 7 | 0 | 0 | 0 | 35398 | 0 | 5 | 6 | 37907 |
| *nen* | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 164 | 0 | 0 | 167 |
| *ns* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 152 | 0 | 152 |
| *s* | 1350 | 0 | 37 | 76 | 13 | 103 | 138 | 18 | 1 | 0 | 85320 | 87056 |
| **Summe** | 246597 | 5339 | 2294 | 9571 | 1114 | 2326 | 3708 | 41064 | 180 | 157 | 87138 | 399488 |

Tab. 8: Konfusionsmatrix vorhergesagte vs. tatsächliche Fugenelemente

In der Diagonalen von Tabelle 8 ist erwartungsgemäß ersichtlich, dass in den meisten Fällen das vorhergesagte auch dem tatsächlichen Fugenelement entspricht.[10] Es sind aber auch die kritischsten Verwechslungen sichtbar: So wird für eine Null-Fuge oder eine *e*-Tilgung des Erstglieds oft stattdessen eine *n*-Fuge oder statt einer Null-Fuge fälschlicherweise eine *s*-Fuge vorausgesagt.[11] Und umgekehrt sind vorausgesagte Null-Fugen in den Daten oft *n*-, *s*-, *en*-, *er*- oder *e*-Fugen.

Wie bereits erwähnt, haben wir für die linguistische Interpretation nicht mit dem kompletten Entscheidungsbaum gearbeitet, da dieser viel zu komplex ist, um daraus linguistisch sinnvolle Regeln abzuleiten. Der im Folgenden interpretierte Baum umfasst pro Blatt mindestens 500 Fälle und wurde im Prozess der Interpretation weiter so zurückgeschnitten, dass sich linguistisch beschreibbare Zusammenhänge erkennen ließen.

[10] Für die Konfusionsmatrix wurden nur die 399488 Fälle betrachtet, in denen das Fugenelement nicht variabel ist.

[11] Die Voraussage der *n*-Fuge statt der korrekten *e*-Tilgung des Erstglieds war bereits oben in Tabelle 7 sichtbar, wo wir feststellten, dass die Tilgung von *e* generell schlecht vorausgesagt wird. Das statistische Modell würde also z.B. im Fall von *Mietwohnung* aufgrund der linguistischen Merkmale der Glieder eher davon ausgehen, dass es *Mietenwohnung* heißen müsste.

Das hat zur Folge, dass der so zurück gestutzte Baum auch nicht mehr 95% der Komposita voraussagen kann, sondern – wie weiter unten detailliert ausgeführt werden wird – etwa 75%.

## 5.7 Interpretation

Durch die geschilderte Vorgehensweise wurde ein Entscheidungsbaum modelliert, der es ermöglicht, Vorhersagen über die Wahl des Fugenelements in nominalen Komposita zu treffen (vgl. Abbildungen 2 und 5). Das auf diese Weise entstandene Modell wird im Folgenden beschrieben und interpretiert sowie in Bezug zu existierenden Hypothesen über die Verteilung von Fugenelementen gesetzt.

Welche Voraussagen trifft unser Modell also für das Auftreten von Fugenelementen, bzw. lässt sich überhaupt eine Fugen-Systematik erkennen? Am obersten Knoten unseres Baummodells wird zunächst spezifiziert, ob es sich beim letzten Laut des Erstglieds um einen Konsonanten oder um einen Vokal handelt. Jede (mitunter komplexe) Bedingung für das Auftreten eines bestimmten Fugenelements beginnt also mit der Spezifizierung des letzten Erstgliedlauts („letzter Laut = Vokal" oder „letzter Laut = Konsonant"). Allerdings handelt es sich im Entscheidungsbaum bei keinem der beiden Fälle um einen terminalen Knoten – zur Vorhersage der Fuge ist jeweils die Bestimmung zusätzlicher Merkmale des Erstglieds notwendig. Das Zweitglied hingegen spielt in unserem Modell so gut wie keine Rolle, da es nur an drei Knoten im Baum, die sich hierarchisch gesehen sehr weit unten befinden, spezifiziert werden muss. Die Merkmale des Zweitglieds sind also nicht nur quantitativ von untergeordneter Bedeutung, sondern kommen wenn überhaupt erst dann ins Spiel, wenn bereits zahlreiche andere Merkmale des Erstglieds festgelegt wurden. Somit wird durch unseren Entscheidungsbaum prinzipiell bestätigt, dass „das Erstglied [...] das Fugenelement [bestimmt]" und Fugenelemente „so und nicht anders [...] systematisch zu fassen sind" (Fuhrhop 1996: 206). Allerdings werden in unserem Modell keine semantischen Eigenschaften des Zweitglieds berücksichtigt.[12]

[12] Nübling/Szczepaniak etwa stellen in ihrer Untersuchung fest, dass nur Zweitglieder mit „ausgeprägter Verbalität" einen Einfluss auf die Kompositionsfuge ausüben können: „Insbesondere deverbale Abstraktbildungen wirken durch ihre noch vorhandene Argumentstruktur auf das Erstglied ein und blockieren dabei tendenziell seine Verfugung", z.B. *Stellungnahme* (Nübling/Szczepaniak 2011: 57).

Aus Gründen der Übersichtlichkeit haben wir den Entscheidungsbaum unterhalb der Entscheidung über den Typus des letzten Erstgliedlautes in zwei Bäume aufgeteilt: Abbildung 2 zeigt alle Knoten nach der Entscheidung „letzter Laut = Konsonant" und Abbildung 5 nach der Entscheidung „letzter Laut = Vokal". Zudem haben wir für die Darstellung den Baum auf die im Folgenden tatsächlich diskutierten Knoten beschränkt. Weggefallen sind Äste, die nur wenige unterschiedliche Komposita und insbesondere wenig unterschiedliche Erstglieder voraussagen, sowie Äste, deren Voraussagen nicht in linguistisch sinnvolle Regeln übertragen werden konnten, da sie z.B. viel zu komplex wären.

### 5.7.1 Entscheidungsbaum Teil A: Erstglied mit auslautendem Konsonanten

Zunächst zu den Voraussagen, die für konsonantisch auslautende Erstglieder getroffen werden können (vgl. Teil A des Entscheidungsbaums in Abbildung 2). Der relevante Teil unseres Modells spricht zunächst deutlich gegen eine „auf den ersten Blick fast regellos wirkende Vielfalt" (Ortner et al. 1991: 68), da allein aufgrund der Erstglied-Endung, d.h. aufgrund der „Wortausgänge" (Augst 1975: 85), direkte und recht zuverlässige Voraussagen für das Auftreten von vier unterschiedlichen Fugenelementen (inklusive Null-Fuge) getroffen werden können.[13]

Mit anderen Worten: Unser Modell benötigt für konsonantisch auslautende Ersteinheiten lediglich eine Spezifizierung der Erstgliedendung, um vorherzusagen, ob eine *nen*-, *en*-, *s*- oder Null-Fuge auftritt (vgl. dazu den Ausschnitt A1 des Entscheidungsbaums in Abbildung 3).

Wenn das Erstglied auf *-in*, *-essin* oder *-nerin* endet, tritt – was aufgrund existierender Hypothesen in der Sekundärliteratur nicht überrascht – in ca. 70% aller berücksichtigter Fälle das Element *nen* in der Fuge entsprechender Komposita auf, z.B. *Lehrerinnenverband*, *Prinzessinnenkleid*, *Wöchnerinnenstation*. Es handelt sich in diesen Fällen prinzipiell um eine *en*-Fuge; da der letzte Kon-

[13] Wie bereits in Kapitel 5.4.2 erwähnt, werden in CELEX zum Teil auch „Wortausgänge" als Suffixe ausgezeichnet, die keine Suffixe im morphologisch engen Sinne sind, z.B. *ern* in *modern*. Daher wird auf dieses CELEX-Merkmal im Folgenden mit dem allgemeineren Begriff „Endung" Bezug genommen. Wenn gekennzeichnet werden soll, dass es sich bei einer Endung um ein Suffix im morphologisch engen Sinne handelt, das der Wortbildung dient, (vgl. auch Bußmann 2008:701, „Suffix") wird dies durch die Verwendung des Begriffs „Ableitungssuffix" kenntlich gemacht.

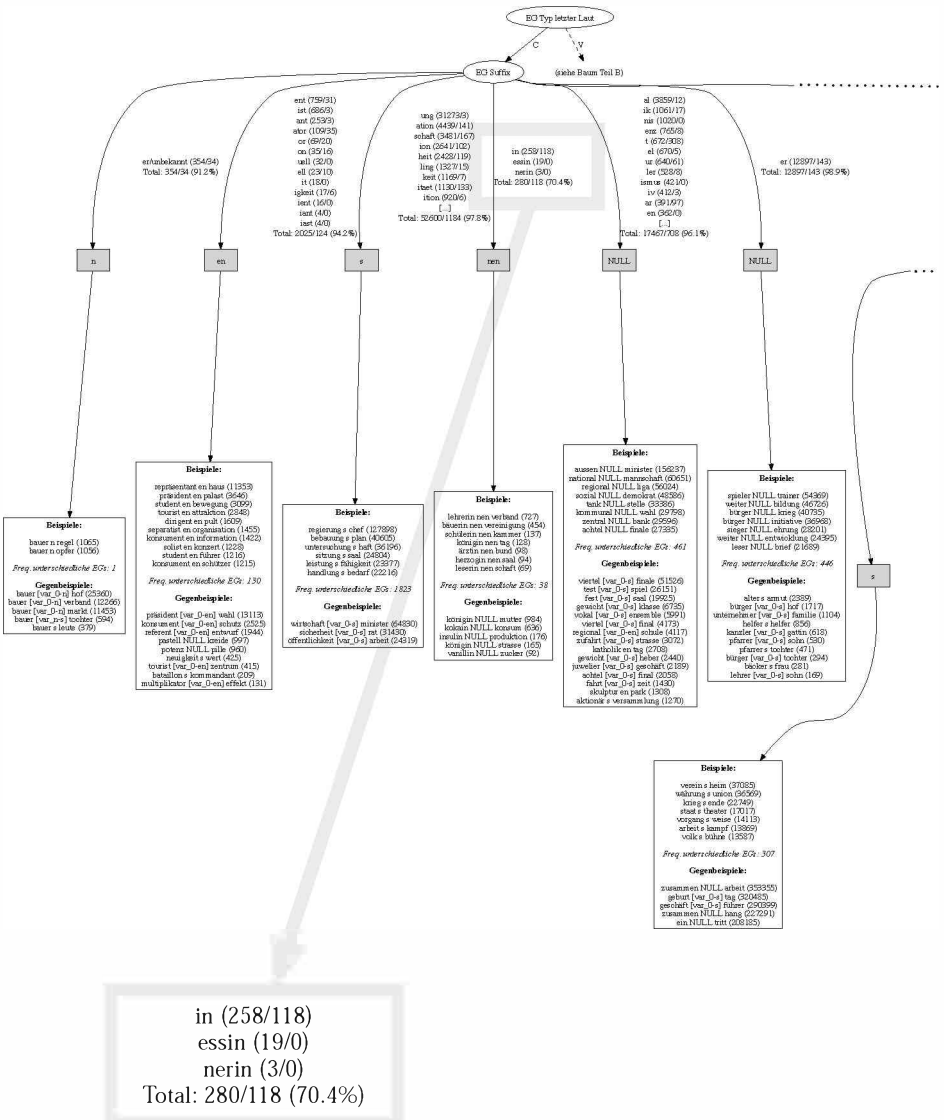
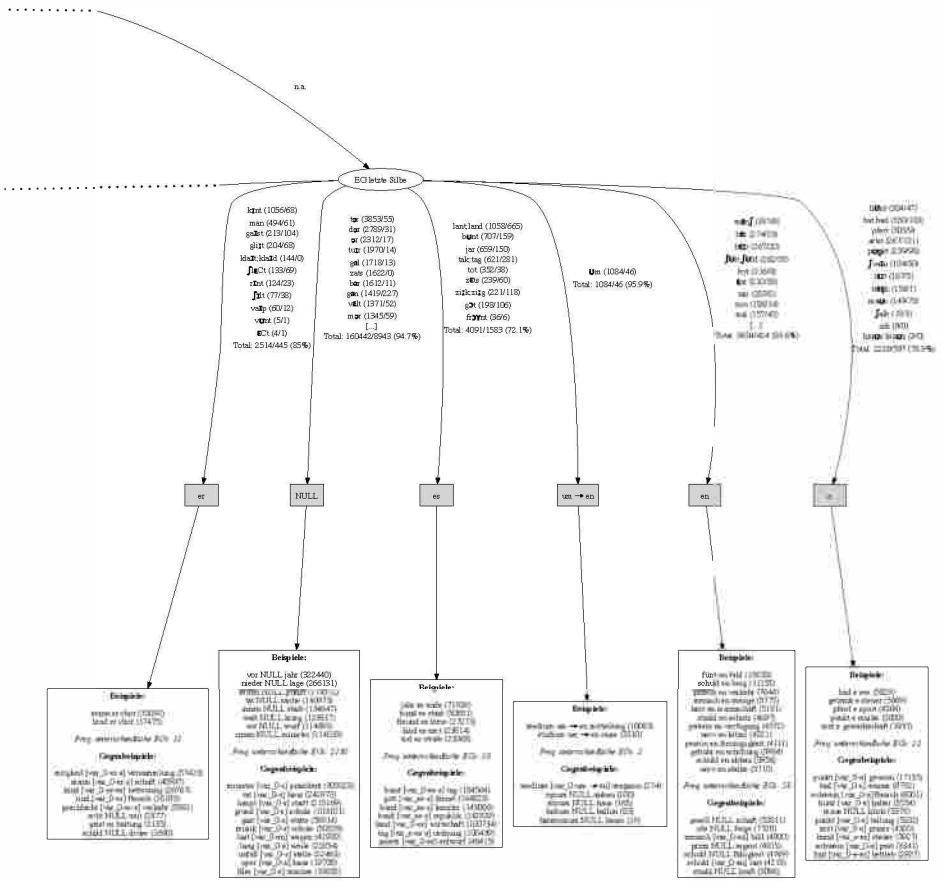



Abb. 2: Entscheidungsbaum Teil A: konsonantisch auslautendes Erstglied; Zahlenwerte in Klammern: Anzahl korrekt und falsch vorausgesagte Komposita, Summenangabe zusätzlich mit prozentualem Anteil der korrekt vorausgesagten Komposita




**Hinweis:** Der hier nur stark verkleinert wiedergegebene Entscheidungsbaum kann unter http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=4731 **als PDF-Dokument in Originalgröße heruntergeladen werden.**

sonant des Erstgliedstamms (*n*) bei der Kompositabildung verdoppelt wird,[14] erfasst unser Modell das zwischen Erst- und Zweitglied auftretende Element allerdings als *nen*.

Besonders aussagekräftig ist die Vorhersage dieser *nen*-Fuge (bzw. dieses modellinhärenten Spezialfalls der *en*-Fuge) für Erstglieder mit dem Ableitungssuffix *-in*, mit dem ausschließlich Feminina gebildet werden (vgl. Altmann 2011: 90). Als mögliche Basen gelten Nomina, die Lebewesen bezeichnen (vgl. *grammis*, „Grammatisches Wörterbuch"). Da es sich auch bei den in unserem Modell enthaltenen Erstgliedern auf *-essin* und *-nerin* durchweg um Feminina handelt, die semantisch gesehen zur Bezeichnung von Lebewesen dienen, kann die folgende Regel formuliert werden:

→ *nen*-Fuge für feminine Nomen auf *-in*, *-essin* oder *-nerin*, die Lebewesen bezeichnen (korrekt: 280; falsch: 118; Abdeckung: 0,07%)[15]

Diese Voraussage deckt sich prinzipiell mit Thesen aus der Sekundärliteratur, so z.B. mit einem Teilergebnis von Ortner et al. zum Auftreten der *en*-Fuge: „Immer steht die Fuge nur bei Feminina auf *-in* (*Lehrerinnenmentalität*)" (Ortner et al. 1991: 94).

Eine *en*-Fuge (ohne Verdopplung des letzten Erstglied-Konsonanten) kann in unserem Modell mit 94-prozentiger Wahrscheinlichkeit vorausgesagt werden. Diese Prognose basiert auf einer (Teil-)Datenmenge von 2 025 Komposita mit 130 unterschiedlichen Erstgliedern. Betrachtet man von den Endungen, über die diese Gruppe definiert ist, nur die, zu denen es mindestens 50 Erstglieder gibt – *-ent* (*Konsumenteninformation*), *-ist* (*Touristenattraktion*) *-ant* (*Gratulantenschar*; *Repräsentantenhaus*), *-(at)or* (*Donatorenliste*) – sind hier klare Regularitäten erkennbar: Alle genannten Endungen sind Ableitungssuffixe, die zur Bildung maskuliner Nomina dienen und lateinischen oder griechischen Ursprungs sind (vgl. *grammis*, „Grammatisches Wörterbuch").

[14] Gemäß Donalies (2011: 30) ist das „gedoppelte[...] n" hier als „reine Schreibgewohnheit" zu betrachten.

[15] Die Zahlenwerte geben an, für wie viele der 407 865 Komposita die Regel gilt. Dabei wird die Anzahl der mit dieser Regel korrekt und falsch vorausgesagten Komposita angegeben. Zusätzlich ist mit „Abdeckung" angegeben, wie viel Prozent der 407 865 Komposita über diese Regel abgedeckt werden. Allerdings ist Folgendes zu berücksichtigen: Die aus dem Entscheidungsbaum abstrahierte linguistische Regel muss nicht unbedingt vollständig mit der Regel im Entscheidungsbaum übereinstimmen. Bei manueller Durchsicht würde man Komposita finden, auf die die Regel des Entscheidungsbaums zwar zutrifft, nicht jedoch die davon abstrahierte linguistische Regel. In den angegebenen Zahlen sind diese Fälle jedoch nicht berücksichtigt.

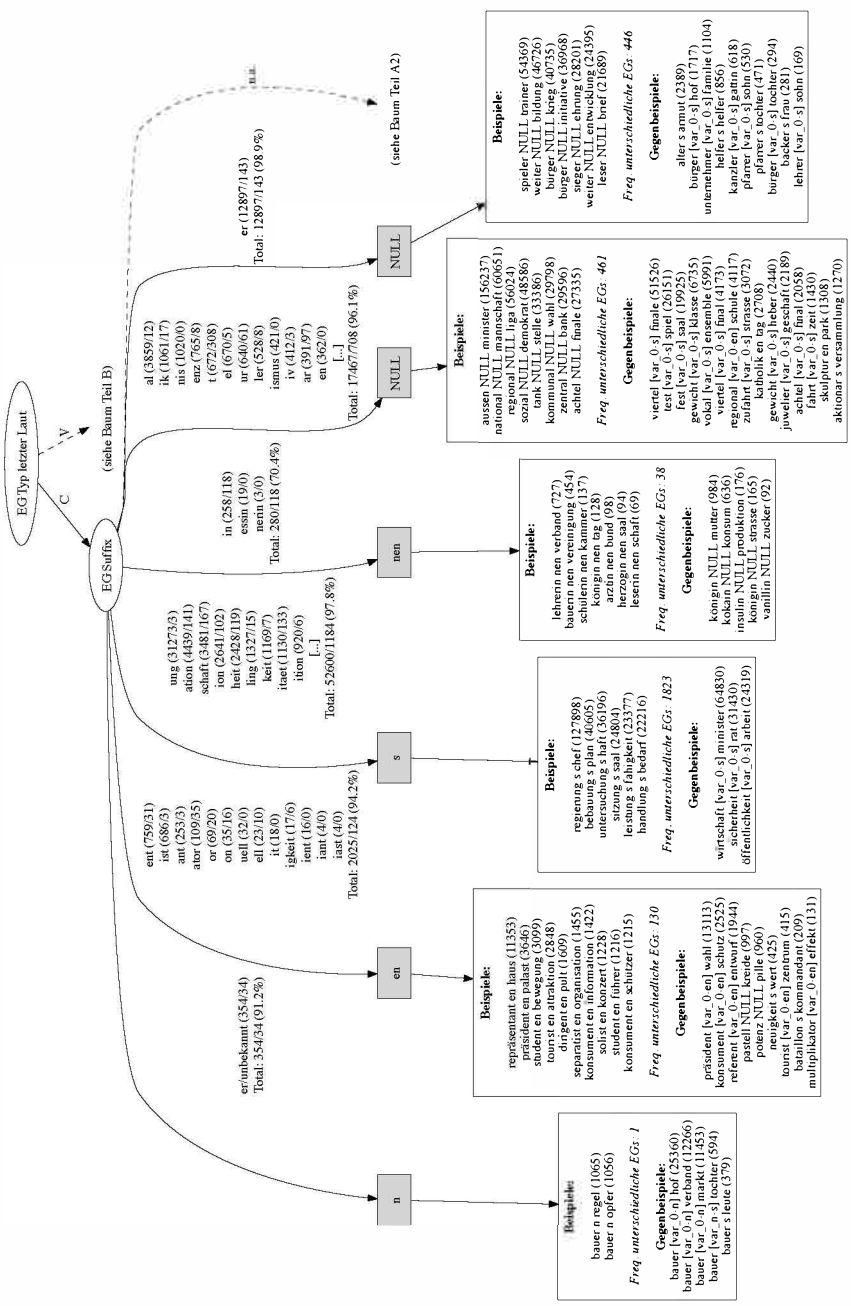



Abb. 3: Entscheidungsbaum Teil A1

→ *en*-Fuge für maskuline Nomina auf *-ent*, *-ist*, *-ant*, *-or* oder *-(at)or* (korrekt: 2025; falsch: 124; Abdeckung: 0,5%)[16]

Ausschlaggebend scheinen innerhalb dieser Gruppe aber weniger die konkreten Ableitungssuffixe als das maskuline Genus bzw. das Flexionsparadigma der suffigierten Erstglieder zu sein: Es handelt sich bei den entsprechenden Erstgliedern aus unserem Modell überwiegend um schwach flektierende Maskulina (*-en* im Genitiv Singular und im Nominativ Plural), die zudem über die Gemeinsamkeit „Bezeichnung von Personen" verfügen, so dass eine eindeutige Übereinstimmung mit Lohde besteht: „Nach einigen Maskulina (im Genitiv Singular/Plural *-en*), die Personen [...] benennen, wird *-en* gesetzt. Hierher gehören auch Nomen mit Fremdsuffixen (v.a. *-ant, -ent, -ist* und *-or*)"[17] (Lohde 2006: 24).

Aufgrund der Erstgliedendung am sichersten vorhersagbar sind die *s*-Fuge (ca. 98% der Fälle) und die Null-Fuge (ca. 96% der Fälle). Zur Voraussage beider Fugen definiert unser Modell jeweils eine sehr große Gruppe verschiedener Erstgliedendungen. Hier stellt sich die Frage, ob für diese beiden Großgruppen tatsächlich gemeinsame abstrakte Merkmale gefunden und somit Regeln modelliert werden können, oder ob keine Systematik erkennbar ist und der Aspekt der Konvention somit den Ausschlag für die Wahl des Fugenelements gibt.

Zunächst zu den Erstgliedern, für die unser Modell eine *s*-Fuge voraussagt und die durch das Vorliegen von 52 600 entsprechenden Komposita innerhalb des hier beschriebenen Baumausschnittes die größte Gruppe bilden:[18] Hier ist insofern eine klare Systematik erkennbar, als der Großteil der enthaltenen Ableitungssuffixe zur Bildung femininer Nomen dient, z.B. *-ung* (*Regierungschef*) *-ion* (*Koalitionsvertrag*), *-keit* (*Abhängigkeitsbericht*), *-schaft* (*Gesellschaftstanz*, *Gemeinschaftserlebnis*, *Gewerkschaftsmitglied*), *-heit* (*Hoheitsgebiet*, *Sicherheitsbericht*, *Weisheitszahn*). Nur vier Ableitungssuffixe weichen von dieser

[16] Auf einige weniger frequente Erstglied-Typen der Gruppe, in der *en* als Fugenelement auftritt, trifft dies nicht zu. Dies gilt z.B. für das Ableitungssuffix *-(ig)keit*, mit dem feminine Nomen gebildet werden. Alle abweichenden Erstglieder sind aber nur vereinzelt in dem hier relevanten Datenausschnitt enthalten, so dass die oben getätigte Generalisierung trotz der Abweichungen gerechtfertigt scheint.

[17] Allerdings zeigen Maskulina auf *-or* in der Regel einen Genitiv auf *-s*, z.B. *des Autors*, *des Donators*.

[18] Hier werden nur Endungen berücksichtigt, die Ableitungssuffixe sind und in *grammis* („Grammatisches Wörterbuch") kodifiziert sind.

Systematik ab, da sie zur Bildung maskuliner Nomina dienen: *-ling*,[19] *-at*, *-tum*, *-eur*. Da der Großteil der Suffixe aus der hier thematisierten Gruppe jedoch über das gemeinsame Merkmal „dient zur Bildung femininer Nomen" verbunden ist, kann nicht von einer regellosen, möglicherweise nur auf sprachlichen Konventionen basierenden Verteilung ausgegangen werden, auch wenn aufgrund der angeführten Abweichungen auf der anderen Seite auch nicht von einer 100-prozentigen Regelhaftigkeit gesprochen werden kann:

→ *s*-Fuge für feminine Nomen auf *-ung*, *-ion*, *-keit*, *-heit* oder *-schaft* (korrekt: 52 600; falsch: 1 209 184; Abdeckung: 12,9%)

Diese Beobachtung ist in ähnlicher Form auch in der Sekundärliteratur verbreitet. Es herrscht ein Konsens darüber, dass die unparadigmische *s*-Fuge „regelmäßig nach Suffixen, die feminine Nomen bilden (*-heit*, *-ion*, *-ität*, *-ung*)" (Duden 2005: 723), auftritt.[20]

Für das Auftreten der *s*- und der *en*-Fuge (ohne Verdopplung des letzten Erstglied-Konsonanten) können also recht eindeutige, deutlich voneinander abgrenzbare Bedingungen formuliert werden (feminine suffigierte Nomen vs. maskuline suffigierte Nomen als Erstglied).

Diese Systematik verliert jedoch an Klarheit und Überzeugungskraft, wenn man die Vorhersagen betrachtet, die unser Modell für das Auftreten einer Null-Fuge trifft. Für die Gruppe von Erstgliedendungen, über die der Ast „führt zu einer Null-Fuge" definiert ist, kann keine eindeutige Regel formuliert werden: Die Gruppe enthält Ableitungssuffixe, mit denen Adjektive gebildet werden (*-al*, *Regionalliga*), Ableitungssuffixe, mit denen feminine Nomen gebildet werden (*-ik*, *Anglistikprofessor*) sowie Ableitungssuffixe, mit denen maskuline Nomen gebildet werden (*-ier*, *Agrarierpartei*). Eine Abgrenzung zwischen Null- und *s*-Fuge bzw. *en*-Fuge kann also nicht durch eine abstrakte grammatische Beschreibung der entsprechenden Ableitungssuffixe, sondern im Zweifelsfall nur durch die Anführung der konkreten Erstgliedendungen erfolgen. Es handelt sich dabei um eine Beobachtung, die deutlich macht, dass die Wahl des Fugenelements – bei aller bisher im Entscheidungsbaum nach-

[19] Auch in der Sekundärliteratur wird postuliert, dass aus der Gruppe der maskulinen Suffixe nur *-ling* regelmäßig mit einem Fugen-*s* auftritt (vgl. Fuhrhop 1996: 537).

[20] Vgl. z.B. auch Fuhrhop (1996: 537), Fleischer (1995: 139). Die unparadigmische *s*-Fuge wird in der Sekundärliteratur außerdem für die Gruppe der „mehrsilbigen femininen Substantive mit dem Auslaut *-t*" (vgl. Duden 2005: 723) vorhergesagt.

gewiesenen Regelhaftigkeit – nicht in allen Fällen durch linguistische Regeln erklärbar bzw. vorhersagbar ist.

Auffallend ist jedoch, dass, im Gegensatz zu den anderen Gruppen des hier beschriebenen Baumausschnittes, etwa ein Drittel der Komposita mit Null-Fuge Adjektive oder Adverbien als Erstglied aufweist (z.B. *sozial*). Daher könnte man aufgrund unseres Modells zumindest postulieren, dass die Null-Fuge tendenziell bei der Integration von suffigierten Adjektiven und Adverbien in ein Kompositum auftritt:

→ Null-Fuge (tendenziell) für Adjektive oder Adverbien als Erstglied (Abdekkung: ~1,4%)[21]

In jedem Fall tritt die Null-Fuge innerhalb der Gruppe der Komposita mit suffigiertem Erstglied deutlich seltener auf als die *s*-Fuge (17 467 vs. 52 600 Komposita). Da nominale Komposita mit adjektivischer Ersteinheit insgesamt deutlich seltener sind als solche mit nominaler Ersteinheit (Lohde 2006: 68), scheint die Postulierung einer Tendenz adjektivischer Erstglieder zur Null-Fuge nicht unplausibel. Auch in der Sekundärliteratur wird davon ausgegangen, dass Adjektive bei der Erstgliedbildung im Normalfall ihre Grundform beibehalten (vgl. z.B. Ortner et al. 1991: 57, Fuhrhop 1996: 529, Fleischer/Barz 1995: 138).

Unter Bezugnahme auf Nübling/Szczepaniak (2011: 57) könnte die quantitative Dominanz der *s*-Fuge über die Null-Fuge innerhalb des hier betrachteten Baumausschnittes der suffigierten Erstglieder damit erklärbar sein, dass „derivationelle Komplexität [...] oft schlechte phonologische Wörter [generiert], die dann umso eher verfugt werden. [...] Besonders schlechte phonologische Wörter werden im Deutschen bereits zuverlässig *s*-verfugt [...]. So haben Derivationssuffixe wie *-schaft*, *-heit* oder *-ung* einen unklaren Status zwischen phonologischem Wort und Reduktionssilbe und werden daher durch das Fugen-*s* hervorgehoben [...]".

Sehr eindeutig vorhersagbar ist die Null-Fuge außerdem für Erstglieder mit dem Ableitungssuffix *-er*, das vor allem zur Bildung maskuliner Nomina dient (vgl. *grammis*, „Grammatisches Wörterbuch"), z.B. *Siegerehrung*, *Bürgerinitiative*, *Anführerrolle*, *Ausländerbehörde*. Hinsichtlich seiner grammatischen

[21] Die komplette Regel des Modells, die auch nicht adjektivische und nicht adverbiale Erstglieder enthält, weist die folgenden Zahlen auf: korrekt: 17 467; falsch: 708; Abdeckung: 4,3%. Etwa ein Drittel der Komposita (also etwa 5 800) lässt sich über die oben formulierte Regel (Null-Fuge für adjektivische oder adverbiale Erstglieder) abdecken. Zwei Drittel bleiben aber unerfasst.

Eigenschaften (Bildung maskuliner Nomina) stimmt *-er* mit den zu einer *en*-Fuge führenden Suffixen (*-ent*, *-ist*, *-ant*, *-or* oder *-(at)or*) überein, so dass die Null-Fuge auch in diesem Fall nur über die konkrete Erstgliedendung, und nicht über abstrakte Suffixeigenschaften, vorhergesagt werden kann:

→ Null-Fuge für Erstglieder mit dem Ableitungssuffix *-er*[22] (korrekt: 12897; falsch: 143; 3,2%)

Im Vorhergegangenen wurden Fälle skizziert, in denen allein aufgrund der Erstgliedendung unmittelbar auf die Ausformung der Kompositionsfuge geschlossen werden kann.[23] Die Endung des Erstglieds bildet innerhalb unseres Entscheidungsbaums demnach ein sehr „mächtiges" Kriterium, oder, unter Einnahme einer anderen Perspektive: Für Erstglieder, die in CELEX als suffigiert ausgezeichnet sind, ermöglicht unser Modell recht eindeutige Voraussagen, die ohne komplexe Verkettungen von Kriterien auskommen. Somit bestätigt unsere statistische Herangehensweise die These, dass „die ausgeprägtesten – wenngleich nicht absolut geltenden – Regelungen [...] sich in Abhängigkeit von bestimmten Suffixen [finden]" (Fleischer/Barz 1995: 139).[24]

Etwas komplexer gestalten sich die Vorhersagen unseres Modells für konsonantisch auslautende Erstglieder, die laut CELEX kein Suffix enthalten (vgl. dazu den Ausschnitt A2 des Entscheidungsbaums in Abbildung 4). Die entsprechenden Ersteinheiten müssen in Bezug auf ein weiteres Merkmal, die phonetische Umschrift der letzten Erstgliedsilbe, spezifiziert werden, damit Aussagen über die auftretende Kompositionsfuge getroffen werden können. Im Folgenden wird jedoch noch deutlich werden, dass sich für die im Modell jeweils gruppierten Letztsilben so gut wie keine abstrakten phonologischen Gemeinsamkeiten erkennen lassen, so dass zur Erschließung der Gruppensys-

[22] Auch Fleischer/Barz postulieren für Erstglieder mit dem heimischen Suffix *-er* eine Null-Fuge (Fleischer/Barz 1995: 139). Eine hochfrequente Abweichung von dieser Regel stellen in unserem Modell Komposita mit dem Erstglied *Alter* dar (*s*-Fuge: *Altersarmut*, *Alterssitz*).

[23] Zwei Äste wurden dabei nicht berücksichtigt: (1) Erstglieder, die laut CELEX zwar suffigiert sind, zu deren Endung es in CELEX aber keine Informationen gibt. (2) Erstglieder, die in CELEX einerseits als nicht-suffigiert ausgezeichnet sind und zugleich den Zusatz enthalten, dass sich die Endung nicht in CELEX befindet.

[24] Unser Modell zeigt, dass nicht nur auf der Grundlage von Suffixen im morphologisch engen Sinne, sondern auch aufgrund von „einfachen" Erstgliedendungen zum Teil recht eindeutige Fugenvoraussagen möglich sind.

tematik – sofern vorhanden – auf andere Merkmale zurückgegriffen werden muss.[25]

Die in CELEX nicht-suffigierten Komposita, für die eine *e*-, *er*- oder *es*-Fuge vorausgesagt wird, weisen jeweils nur zehn bis zwölf unterschiedliche Erstglieder auf. Betrachtet man die entsprechenden Gruppen genauer, ist außerdem auffällig, dass die Letztsilben, über die die Gruppen bestimmt sind, fast ausschließlich jeweils nur innerhalb eines Erstglieds auftreten. Beispielsweise kommt die Silbe *-trɛŋk* nur mit dem Erstglied *Getränk* vor (*Getränkemarkt*). Da weder zwischen den entsprechenden Letztsilben noch zwischen den entsprechenden Erstgliedern abstrakte Gemeinsamkeiten vorliegen, scheint die Voraussage einer *e*-, *er*- oder *es*-Fuge innerhalb unseres Modells weniger mit den Eigenschaften der letzten Erstgliedsilbe als mit einem konkreten Erstgliedlexem zusammenzuhängen. Die Wirksamkeit solcher lexikalischer Konventionen ist besonders im Fall von einsilbigen Erstgliedern wie z.B. *Freund* augenfällig, da hier durch die Angabe der Letztsilbe *-frɔʏnt* das gesamte Erstgliedlexem abgebildet wird (*es*-Fuge: *Freundeskreis*).[26] Die Tatsache, dass die Erstglieder der Komposita mit *e*-, *er*- oder *es*-Fuge überwiegend einsilbig sind und durch die Angabe der Letztsilben somit überwiegend eine Abbildung des gesamten Erstglieds erfolgt, liefert positive Evidenz für die Wirksamkeit idiosynkratischer lexikalischer Konventionen an dieser Stelle im Entscheidungsbaum.[27]

Für das Auftreten von *s*-, Null- und *en*-Fuge werden in unserem Modell jeweils Gruppen mit einer deutlich höheren Anzahl unterschiedlicher Erstglieder (ca. 60 bis 2200) definiert. Zudem finden sich die Letztsilben, über die die Gruppen bestimmt sind, nicht durchgängig nur in einem einzelnen Erstglied, sondern treten (zumindest zum Teil) innerhalb einer Vielzahl unterschiedlicher Erstglieder auf. Allgemein lassen sich für die Gruppe der Komposita, für die im Entscheidungsbaum eine *s*-, Null- oder *en*-Fuge vorhergesagt wird, eher Regularitäten – wenn auch keine phonologisch motivierten – erkennen als für das Auftreten von *e*-, *er*- oder *es*-Fuge.

[25] Nur weil die Letztsilben im Entscheidungsbaum in phonetischer Umschrift codiert wurden, heißt das nicht, dass der Algorithmus hier eine phonologische Regel gefunden hat. Vielmehr ist die Eigenschaft „Letztsilbe" an diesem Punkt unter den berücksichtigten Eigenschaften nur die beste Möglichkeit, die Erstglieder zu gruppieren.

[26] Auch in der Sekundärliteratur wird angenommen, dass das Fugenelement *es* überwiegend bei einsilbigen Erstgliedern auftritt (z.B. Duden 2005: 723).

[27] An dieser Stelle wird darauf verzichtet, die Silben (bzw. Erstglieder) aufzulisten, die jeweils zu einer *e*-, *er*- bzw. *es*-Fuge führen. Diese können Abbildung 4 entnommen werden.

Betrachtet man die in CELEX nicht-suffigierten Erstglieder mit *s*-Fugen-Voraussage, können, zumindest tendenziell, zwei linguistische Regeln formuliert werden. Erstens scheint ein Zusammenhang zwischen dem Auftreten einer *s*-Fuge und dem Flexionsparadigma des Erstglieds zu bestehen, da *s* für die Mehrheit der enthaltenen Erstglieder paradigmisch ist, z.B. *Vereinsheim*, *Geschäftsbereich*, *Kriegsende*:

→ (1) *s*-Fuge (tendenziell) für nicht-suffigierte Erstglieder, für die die *s*-Fuge paradigmisch ist[28]

Jedoch trifft diese abstrakte Beschreibung nicht auf alle Erstglieder zu, für die in diesem Baumausschnitt eine *s*-Fuge vorausgesagt wird. Beispielsweise handelt es sich beim Erstglied *Weihnacht*, das in unseren Daten in 508 Komposita vorkommt, um eine sehr frequente Ausnahme im Sinne einer Bildung, die von der in (1) formulierten Eigenschaft nicht erfasst wird, aber dennoch mit einer *s*-Fuge auftritt.

Unter Berücksichtigung der Beobachtung, dass die Gruppe der suffixlosen Komposita mit *s*-Fuge fast ausschließlich mehrsilbige und/oder derivationell komplexe Erstglieder[29] aufweist (z.B. *Arbeitskampf*, *Lebensweise*, *Abzugsabkommen*, *Vereinsheim*), lässt sich zudem die folgende Regel ansetzen:

→ (2) *s*-Fuge (tendenziell) für nicht-suffigierte Erstglieder, die mehrsilbig und/oder derivationell komplex sind (für 1 und 2 zusammen gilt: korrekt: 32 210; falsch: 6 219; Abdeckung: 7,9%)

Durch die in (2) formulierte Regel bestätigt sich einerseits die vorherrschende Meinung, dass die *s*-Fuge „bevorzugt nach mehrsilbigen [Erstgliedern]" (Fuhrhop 1996: 537) auftritt. Andererseits liefert (2) positive Evidenz für die in Verbindung mit suffigierten Erstgliedern bereits angeführte These vom Zusammenhang zwischen *s*-Fuge und derivationell komplexen Erstgliedern bzw. zwischen *s*-Fuge und schlechten phonologischen Wörtern (Nübling/Szczepaniak 2011: 57).

Unser Modell bestätigt an dieser Stelle im Baum allerdings nicht, dass vor allem mehrsilbige Nomen, die auf *-t* enden (vgl. z.B. Fuhrhop 1996: 537) im Kompositum eine *s*-Fuge nehmen, auch wenn der Entscheidungsbaum entsprechende Fälle enthält, z.B. *Ankunftsabend*.

[28] „Da ein Erstglied mit paradigmischem *-s-* niemals einer Pluralform entspricht, kann es nur der GenSg-Form homonym sein" (Fuhrhop 1996: 535).

[29] Die Erstglieder sind zwar in CELEX als nicht-suffigiert ausgezeichnet, können aber über ein Präfix verfügen, das dann zu derivationeller Komplexität führt.

Nun zu der Gruppe von nicht-suffigierten Erstgliedern, für die unser Modell eine *en*-Fuge voraussagt: Die Erstglied-Letztsilben, über die die Gruppe definiert ist, sind nicht sehr frequent – mit einem Vorkommen in vier unterschiedlichen Erstgliedern stellt *-graf* die in dieser Hinsicht frequenteste Silbe dar (*Markgraf*, *Fotograf*, *Telegraf*, *Graf*). Auch hier scheint aber ohnehin – wie auch schon im Fall der *s*-Fuge – keine linguistische Regel formulierbar, die auf den jeweiligen Letztsilben oder deren Lautung basiert. Vielmehr lässt sich auch hier ein Zusammenhang zwischen dem auftretenden Fugenelement und dem Flexionsparadigma des Erstglieds feststellen: Es handelt sich bei den entsprechenden Erstgliedern aus unserem Modell überwiegend um schwache Maskulina (*-en* im Genitiv Singular und im Nominativ Plural), so dass *en* hier als paradigmisches Fugenelement auftritt, z.B. *Menschenmenge*, *Herrenabend*, *Bärenanhänger*:

→ (1) *en*-Fuge für nicht-suffigierte, schwach flektierende Maskulina

Diese Regel deckt sich insofern mit einer anderen Stelle im Modell, als in Verbindung mit suffigierten Erstgliedern für schwach flektierende Maskulina, die der Bezeichnung von Personen dienen, bereits eine Tendenz zur *en*-Fuge festgestellt wurde.[30]

Die Gruppe der nicht-suffigierten Erstglieder mit *en*-Fuge enthält jedoch auch feminine Nomen, z.B. *Schuldenberg*, *Instanzenweg*, *Personenverkehr*. Mit Fuhrhop könnte man das Auftreten des Fugenelements *en* in solchen Zusammensetzungen mit femininem Erstglied auf das Vorliegen positiver Pluralmarkierungen zurückführen (vgl. Fuhrhop 1996: 541):

→ (2) *en*-Fuge für nicht-suffigierte, feminine Erstglieder, die tatsächlich einen Plural ausdrücken (für 1 und 2 zusammen gilt: 3 634 korrekt; 424 falsch; Abdeckung 0,9%)

Schließlich soll noch auf die Gruppe nicht-suffigierter Erstglieder eingegangen werden, für die unser Modell eine Null-Fuge voraussagt. Betrachtet man innerhalb dieser Gruppe die frequentesten Letztsilben genauer – d.h. die Letztsilben, die in vielen unterschiedlichen Erstgliedern auftreten – fällt auf, dass die entsprechenden Erstglieder größtenteils auf „*e* + Konsonant" (genauer: nach CELEX auf „Schwa + Konsonant") enden, z.B. *Abendkasse*, *Jugendarbeit*,

[30] Diese Tendenz schwacher Maskulina zur (*e*)*n*-Fuge wird auch in Verbindung mit der Fugenmodellierung für vokalisch auslautende Erstglieder wieder auftreten.

*Titelverteidiger*, *Körperverletzung*, *Winterpause*, *Steuerzahler*.[31] Folglich kann in Verbindung mit den Letztsilben zumindest für die Null-Fuge eine Regel formuliert werden, in der lautliche Aspekte immerhin eine minimale Rolle spielen:

→ (1) Null-Fuge (tendenziell) für nicht-suffigierte Erstglieder, die auf „*e* + Konsonant" enden

Diese Regel verliert jedoch an Klarheit bzw. Überzeugungskraft, wenn man anstelle der frequentesten Letztsilben die frequentesten Erstglieder als Ausgangspunkt nimmt. Aus dieser Perspektive ist vielmehr auffallend, dass die Gruppe der Komposita mit Null-Fuge vor allem dadurch bestimmt ist, dass es sich überwiegend um Simplizia handelt, z.B. *Weltkrieg*, *Abendabitur*:

→ (2) Null-Fuge (tendenziell) für simplizische Erstglieder (für 1 und 2 zusammen gilt: korrekt: 160 442; 8 943 falsch; Abdeckung 39,3%)

Zum Abschluss der Betrachtung von Erstgliedern, die in CELEX als endungslos ausgezeichnet sind, sei noch darauf verwiesen, dass die Voraussagen unseres Modells hier zwar prinzipiell in linguistische Regeln überführt werden können, es sich aber um eine Stelle im Entscheidungsbaum handelt, an der sich die dem Modell zugrundeliegende Systematik nur schwer erschließen lässt. Trotz der angeführten linguistischen Regeln sollte nicht ausgeschlossen werden, dass lexikalische Konventionen an dieser Stelle im Modell eine „mächtige" Rolle spielen.

### 5.7.2 Entscheidungsbaum Teil B: Erstglied mit auslautendem Vokal

Während in unserem Modell für konsonantisch auslautende, suffigierte Erstglieder aufgrund ihrer Endung unmittelbar auf die Gestaltung der Kompositionsfuge geschlossen werden kann und auch die Voraussagen für konsonantisch auslautende, nicht-suffigierte Erstglieder ohne komplizierte Regelverkettungen auskommen, gestaltet sich die Vorhersage von Fugenelementen für vokalisch auslautende Erstglieder grundsätzlich etwas komplexer (vgl. den Teil B des Entscheidungsbaums in Abbildung 5). Zunächst ist im Entscheidungsbaum ausschlaggebend, ob die letzte Silbe des Erstglieds betont oder unbetont ist. Die Spezifizierung dieses Kriteriums führt jedoch nicht zu terminalen

[31] Alle genannten Fälle werden in CELEX mit Schwa vor dem jeweiligen Konsonanten codiert; allerdings gibt es einen großen Interpretationsspielraum und Fälle wie *Körper* könnten z.B. auch mit einem vokalisierten *r* ([*kœʁpɐ*]) codiert werden.

Knoten. Vielmehr müssen weitere Merkmale spezifiziert werden, um Voraussagen über die entsprechenden Kompositionsfugen treffen zu können.

Für vokalisch auslautende Erstglieder mit betonter Letztsilbe muss zur Fugen-Vorhersage die phonetische Umschrift der letzten Silbe spezifiziert werden (vgl. den Ausschnitt B1 des Entscheidungsbaums in Abbildung 6). Für diese Gruppe von Erstgliedern fällt bereits unabhängig von der Berücksichtigung lautlicher Details auf, dass deren Integration in ein Kompositum überwiegend ohne Fugenelement erfolgt: Unsere Daten enthalten insgesamt 14867 entsprechende Komposita mit Null-Fuge (z.B. *Polizeisprecher, Neuordnung*), während nur 963 Zusammensetzungen vorliegen, in denen ein Fugenelement realisiert wird. Zusätzlich zu dieser vergleichsweise geringen Gesamtfrequenz muss hier noch berücksichtigt werden, dass sich innerhalb dieser 963 Komposita mit Fugenelement nur vier[32] unterschiedliche Erstglieder befinden (= *Idee*,[33] *treu*, *Frau*, *Papagei*),[34] während innerhalb der Komposita mit Null-Fuge 288 unterschiedliche Erstglieder vorliegen.

Da die Gruppe der Komposita mit Null-Fuge in unserem Modell über eine Vielzahl von Erstglied-Merkmalen definiert ist, wird an dieser Stelle darauf verzichtet, die einzelnen Merkmale anzuführen. Vielmehr kann die aus dem Entscheidungsbaum ableitbare Regel für das Auftreten einer Null-Fuge am einfachsten in Form der folgenden negativen Definition formuliert werden:

→ Null-Fuge für Erstglieder mit betonter Letztsilbe, deren letzte Silbe *nicht* die Lautung *-de, -laʊ, -trɔʏ, -fraʊ, -gaɪ* oder *-aɪ* hat (korrekt: 14867; falsch: 444; Abdeckung: 3,7%).

Ist die letzte Silbe vokalisch auslautender Erstglieder hingegen unbetont, muss in unserem Modell zur Vorhersage der Kompositionsfuge die phonetische Umschrift des letzten Lautes, nicht die der letzten Silbe, des Erstglieds bestimmt werden (vgl. dazu den Ausschnitt B2 des Entscheidungsbaums in Abbildung 7). Endet ein entsprechendes Erstglied nicht auf Schwa, tritt ebenfalls kein Fugenelement auf, z.B. *Kilogramm, Klimaschutz, Taxifahrer, Abbauantrag* usw.

[32] Das Modell fand noch zwei weitere Erstglieder, die jedoch auf eine falsche morphologische Analyse durch CELEX zurück gehen: *Ei* (in Abbildung 6 nicht dargestellt), z.B. in *Eiscafé*, das als *Ei-s-café*, und *Lau* (siehe Abbildung 6 unter *e*-Fuge), z.B. in *Lauegymnasium*, das als *Lau-e-gymnasium* analysiert wurde.

[33] Bildungen auf *-ee* unterliegen laut Fleischer/Barz bezüglich der Fugengestaltung stärkeren Schwankungen (vgl. Fleischer/Barz 1995: 139).

[34] In Abbildung 6 nicht als Beispiel aufgeführt, gehört jedoch mit dem Gesamtkomplex *Papageienart* zur *en*-Fuge.

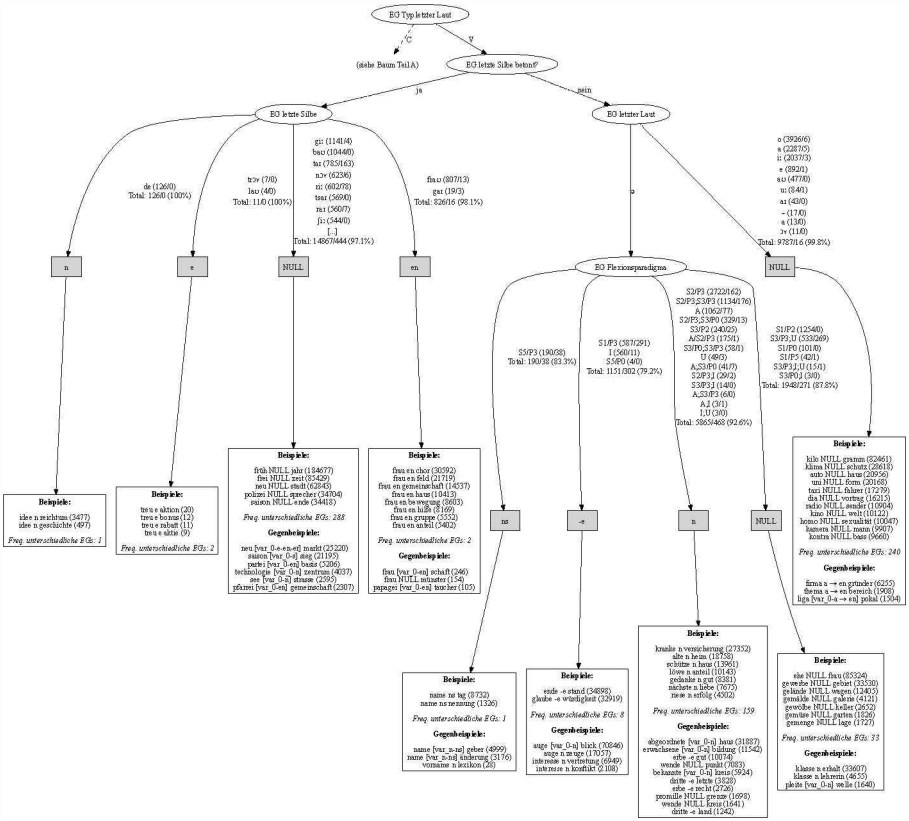



**Abb. 4: Entscheidungsbaum Teil B: mit auslautendem Vokal des Erstglieds; Zahlenwerte in Klammern: Anzahl korrekt und falsch vorausgesagte Komposita, Summenangabe zusätzlich mit prozentualem Anteil der korrekt vorausgesagten Komposita**

→ Null-Fuge für Erstglieder mit unbetonter Letztsilbe, deren letzter Laut nicht Schwa ist (korrekt: 9 787; falsch: 16; Abdeckung: 2,4%).[35]

Endet das Erstglied hingegen mit einem Schwa, muss zur Vorhersage der Kompositionsfuge das Kriterium „Flexionsparadigma des Erstglieds“[36] berücksichtigt werden – die Flexionsklasse gilt neben Wortart und Lautstruktur

[35] Recht frequente Ausnahmen von dieser Regel sind Komposita mit dem Erstglied *Firma* (z.B. *Firmengründer*) und Komposita mit dem Erstglied *Thema* (z.B. *Themenkomplex*).

[36] Die folgenden Ausführungen orientieren sich an der CELEX-Klassifikation, für die bereits auf Abweichungen von anderen üblichen Flexionsparadigmen hingewiesen wurde (vgl. dazu Kapitel 5.4.2).

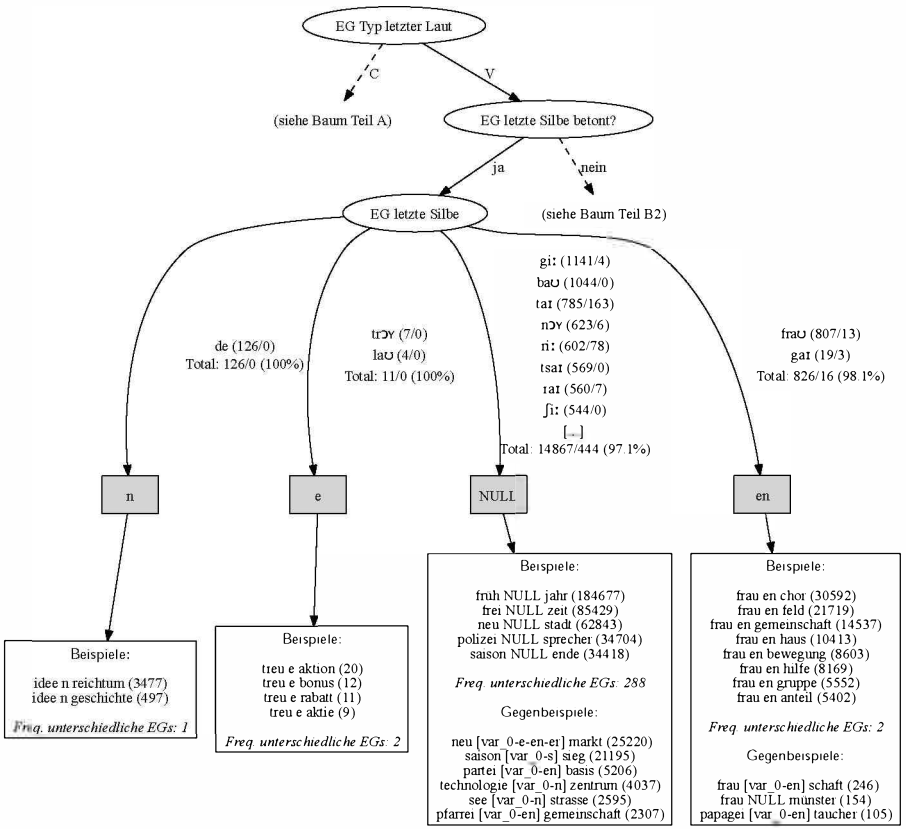



**Abb. 5: Entscheidungsbaum Teil B1**

allgemein als einer der Faktoren des Erstglieds, die Einfluss auf die Gestaltung der Fuge haben (z.B. Lohde 2006: 22).

Als eine Art Zwischenfazit kann bereits an dieser Stelle im Entscheidungsbaum konstatiert werden, dass das Flexionsparadigma des Erstglieds bei konsonantisch auslautenden Erstgliedern so gut wie keine Rolle spielt,[37] während es bei vokalisch auslautenden Erstgliedern deutlich relevanter ist und sich in der Hierarchie des Entscheidungsbaums relativ weit oben – wenn auch unterhalb einiger lautlicher Merkmale – befindet.

[37] Bei konsonantisch auslautenden Erstgliedern spielt das Merkmal „Flexionsparadigma" nur eine Rolle, wenn man Erstglieder berücksichtigt, deren Suffix nicht in CELEX enthalten ist. Wie bereits erwähnt, wurden solche Erstglieder hier aber nicht betrachtet, da nicht transparent genug ist, um welche Art von Suffigierung es sich dabei handelt.

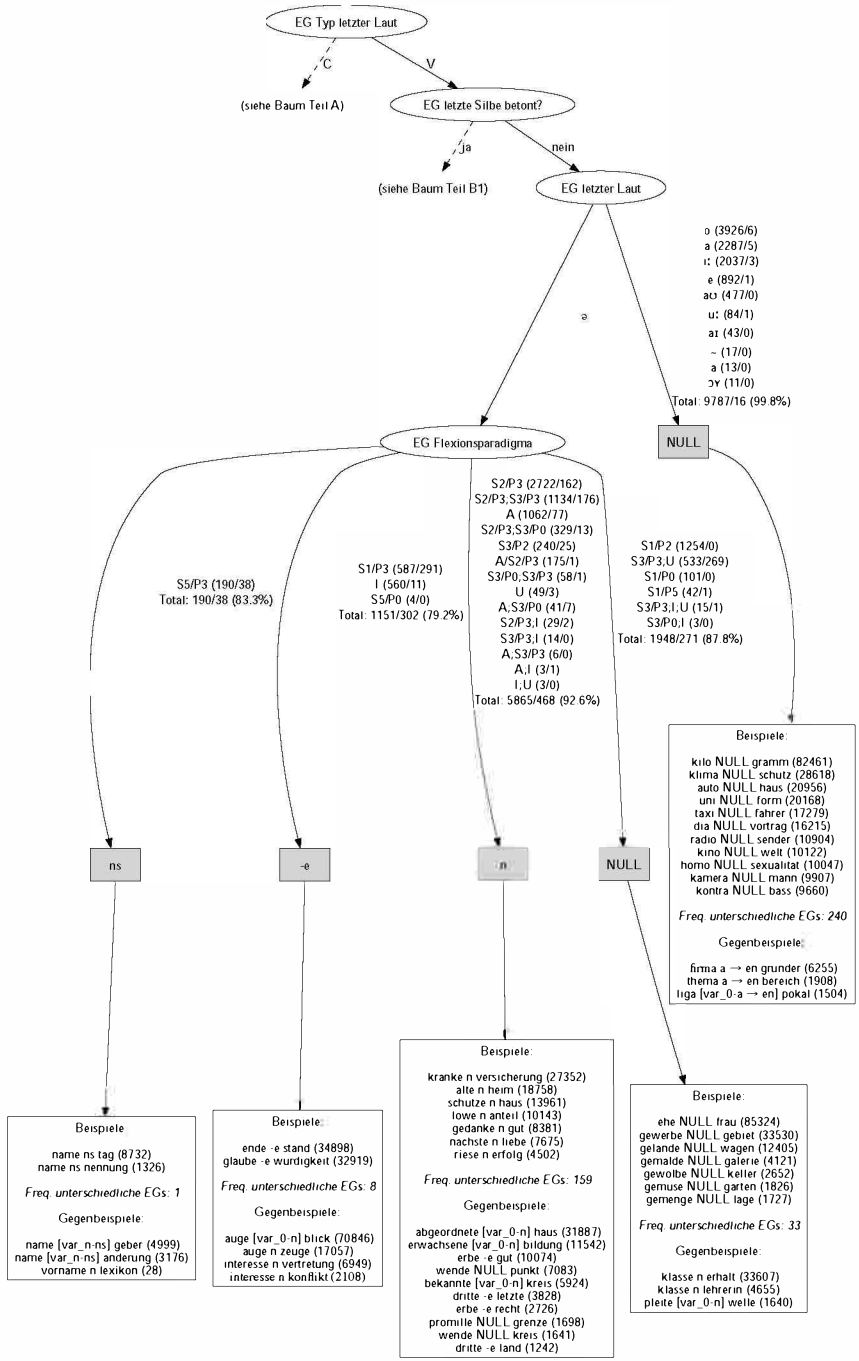



**Abb. 6: Entscheidungsbaum Teil B2**

Welche Vorhersagen können durch die Spezifizierung des Flexionsparadigmas für Ersteinheiten auf Schwa mit unbetonter Letztsilbe in unserem Modell getroffen werden? Zunächst zu den vier Fällen, in denen anhand der zusätzlichen Spezifizierung des Flexionsparadigmas direkt auf die Kompositionsfuge geschlossen werden kann.

Eine Null-Fuge tritt mit einer Wahrscheinlichkeit von ca. 88% mit Erstgliedern auf, die wie folgt flektieren:[38]

(1) -(*e*)*s* im Genitiv-Singular/endungslos im Nominativ Plural (nominales Flexionsparadigma) → (*e*)*s*/- → starke Nomen
   - Beispiel aus unserem Modell: *Gewerbegebiet*, *Gemäldegalerie*

(2) -(*e*)*s* im Genitiv Singular/nur im Singular gebräuchlich (nominales Flexionsparadigma) → (*e*)*s*/Singularia Tantum → starke Nomen
   - Beispiel aus unserem Modell (die Beispiele gehen nur auf zwei unterschiedliche Erstglieder zurück): *Prestigedenken, Karatemeister*

→ Null-Fuge für nominale Erstglieder auf Schwa, die im Genitiv Singular auf -(*e*)*s* enden und im Nominativ Plural endungslos oder nur im Singular gebräuchlich sind (korrekt: 1 355; falsch: 0; Abdeckung: 0,33%).[39]

Prinzipiell bestätigt unser Modell damit in der Sekundärliteratur formulierte Regularitäten: Gemäß Fuhrhop (1996: 542) wird Schwa u.a. immer dann beibehalten – in unserer Terminologie entspricht dies dem Auftreten einer Null-Fuge – wenn das Erstglied keinen Plural kennt, keinen *n*-Plural hat oder der Plural ungewöhnlich ist. Die erstgenannte Bedingung Fuhrhops trifft auch in unserem Entscheidungsbaum eindeutig zu, während der Bedingung „kein *n*-Plural" in unserem Modell zumindest durch das Auftreten von den Erstgliedern *Ehe* und *Pleite* mit einer Null-Fuge widersprochen wird. Die in (2) formulierte Bedingung für das Auftreten der Null-Fuge stimmt eindeutig mit einer These von Fleischer/Barz überein: „-0- steht bei Singulariatantum" (Fleischer/Barz 1995: 139).

[38] Hier werden nur Flexionsparadigmen berücksichtigt, zu denen in unseren Daten mindestens 50 Exemplare vorliegen.

[39] Eine weitere Null-Fugen-Gruppe mit mehr als 50 Exemplaren (CELEX-Flexionsparadigma S3/P3;U) wird von dieser Regel nicht erfasst. Sie enthält jedoch nur zwei unterschiedliche Erstglieder (*Ehe*, *Pleite*).

Neben der Null-Fuge sagt unser Modell für die Kompositabildung bestimmter Erstglieder auf Schwa außerdem die Tilgung des Letztlauts voraus. Hier wird ersichtlich, dass für die „Problematik der Kompositionsfuge" auch die „Tilgung bestimmter Elemente [...] eine Rolle" (Fleischer/Barz 1995: 136) spielt.

Es handelt sich dabei in unserem Modell aber um eine deutlich weniger aussagekräftige Prognose als bei der Vorhersage der Null-Fuge, da die entsprechende Datengrundlage nur acht unterschiedliche Erstglieder (aber insgesamt 1 151 Komposita) aufweist: Die in dieser Gruppe enthaltenen Lexeme sind zum einen Nomina, die im Genitiv Singular *-(e)s* und im Nominativ Plural *-n* nehmen (Flexionsklasse *(e)s/n*) oder im Genitiv Singular *-ns* nehmen und nur im Singular gebräuchlich sind (Flexionsklasse: *ns*/Singularia Tantum). Zum anderen wird die Schwa-Tilgung für 560 Erstglieder vorausgesagt, für die in CELEX zwar angegeben wird, dass sie flektieren, für die aber kein Paradigma verfügbar ist. Beispiele für entsprechende Erstglieder sind *Ende* (*Endstand*), *Glaube* (*Glaubwürdigkeit*), *zweite* (*Zweitliga*), *dritte* (*Drittmittel*). Auffallend ist hier, dass ein Zusammenhang zwischen Ordinalzahlen als Erstglied und der Schwa-Tilgung zu bestehen scheint:

→ Schwa-Tilgung bei Ordinalzahlen als Erstglied (Abdeckung: < 0,2%)[40]

Die Formulierung einer weiteren Regel, z.B. die Herstellung eines direkten Zusammenhangs zwischen Schwa-Tilgung und Flexionsparadigma, ist hingegen schwierig, zumal die beschriebene Gruppe nur acht unterschiedliche Erstglieder enthält, darunter allein fünf Ordinalzahlen. Die Schwa-Tilgung wird in unserem Modell u.a. für das Nomen *Ende* (z.B. *Endstation*) vorausgesagt, für das auch in der Sekundärliteratur der „regelmäßig[e] Ausfall des stammauslautenden *-e*" (Fleischer/Barz 1995:71) postuliert wird.[41]

Das Auftreten der *n*-Fuge für auf Schwa auslautende Erstglieder kann am ehesten durch eine negative Definition modelliert werden, da diese Fuge in unserem Modell über eine Vielzahl möglicher Flexionsparadigmen bestimmt wird: Bei Erstgliedern mit anderen Flexionsparadigmen (als bei den Erstein-

[40] Der Entscheidungsbaum weist in dieser Gruppe keine Differenzierung zwischen Ordinalzahlen als Erstgliedern und anderen Erstgliedern auf. 545 der 849 Komposita dieser Gruppe weisen Ordinalzahlen-Erstglieder auf. Die Regel deckt demnach weniger als 0,2% der Komposita ab.

[41] Auch Fuhrhop nennt entsprechende Bildungen, merkt jedoch an, dass die Tilgung des Schwas „weiterhin ungeklärt [bleibt]" (Fuhrhop 1996: 543).

heiten mit Null-, *ns*-Fuge[42] oder Schwa-Tilgung) tritt innerhalb einer Zusammensetzung ein *n* als Fugenelement auf. Unsere Daten enthalten an die 6000 entsprechende Komposita, so dass die *n*-Fuge bei der Kompositabildung mit auf Schwa auslautenden Ersteinheiten (mit unbetonter Letztsilbe) in unserem Modell am üblichsten ist.[43] Auch Fuhrhop bezeichnet es als den „Normalfall", „daß Schwa in der Komposition *-n-* nimmt" (Fuhrhop 1996: 541).

Auch wenn das Auftreten der *n*-Fuge zunächst negativ definiert wurde, sollen die entsprechenden Erstglieder hinsichtlich ihres Flexionsparadigmas noch genauer spezifiziert werden: Es handelt sich überwiegend um schwach flektierende Maskulina – Flexionsparadigma *n/n* – wie z.B. *Kranke* (*Krankenversicherung*), *Schütze* (*Schützenhaus*) oder *Franzose* (*Franzosenduo*). Die auftretende *n*-Fuge ist jeweils paradigmisch.

→ *n*-Fuge für schwach flektierende Maskulina, die auf Schwa auslauten (korrekt: 5865; falsch: 468; Abdeckung: 1,4%).

Diese recht deutlich hervortretende Regelhaftigkeit wurde in ähnlicher Form bereits an einer anderen Stelle im Modell konstatiert: Für suffigierte, schwach flektierende Maskulina, die mit einem Konsonanten enden, wurde eine eindeutige Tendenz zur *en*-Fuge festgestellt (vgl. Interpretation der Erstglieder mit auslautendem Konsonanten). Auch in der Forschungsliteratur wird die Tendenz zur *(e)n*-Fuge in allgemeiner Form, d.h. unabhängig vom Letztlaut, erfasst: „Die Fugenelemente *en* und *n* erscheinen paradigmisch bei schwach flektierenden Maskulina." (Duden 2005: 723). Fuhrhop formuliert die „hinreichende Bedingung", dass „alle schwach flektierenden Maskulina [...] in ihrer flektierten Form im Kompositum auf[treten]" (Fuhrhop 1996: 541) und konkretisiert diese Aussage dann noch dahingehend, dass das Fugenelement *n* nur nach Schwa auftritt (vgl. ebd.: 543).

Neben den vorstehend erläuterten vier terminalen Knoten führt die Spezifizierung des linguistischen Kriteriums „Flexionsparadigma" für die auf Schwa auslautenden Erstglieder in unserem Entscheidungsbaum auch zu nichtterminalen Knoten. Da in der Folge sehr komplexe Regelverkettungen entste-

[42] Die *ns*-Fuge wurde im Vorhergegangenen nicht näher betrachtet, da sie von unserem Modell nicht musterhaft vorausgesagt wird. Sie tritt nur mit einem bestimmten auf Schwa auslautenden Erstglied – *Name* – auf, z.B. N*amensregister, Namenspatrone, Namenstag, Namensnennung. Name* flektiert gemäß des *ns/n*-Paradigmas, so dass das auftretende Fugenelement *ns* der Genitiv-Singular-Form entspricht und somit paradigmisch ist.

[43] Zum Vergleich: Unsere Daten enthalten für auf Schwa auslautende Erstglieder 1948 Komposita mit Null-Fuge, 1151 Komposita mit Schwa-Tilgung und 190 Komposita mit *ns*-Fuge.

hen, deren Erläuterung für die Zwecke des vorliegenden Aufsatzes nicht zielführend ist, wird der entsprechende Teil des Entscheidungsbaums an dieser Stelle nicht weiter beschrieben. Es sei lediglich erwähnt, dass zur Fugen-Vorhersage für auf Schwa auslautende Erstglieder mit bestimmtem Flexionsparadigma eines oder mehrere der folgenden Merkmale aus CELEX spezifiziert werden müssen: morphologischer Status des Erstglied-Präfixes bzw. -Suffixes (wenn kein Präfix vorhanden), Lautung der letzten Erstgliedsilbe, Silbenanzahl des Erstglieds, Position der betonten Silbe innerhalb des Erstglieds, Betontheit der ersten Zweitgliedsilbe, morphologischer Status des Zweitglieds.

Allein aufgrund dieser Andeutungen zum weiteren Verlauf der Modellierung sollte deutlich geworden sein, dass Erstglieder mit dem Letztlaut Schwa innerhalb der Gruppe der vokalisch auslautenden Erstglieder eine deutliche Sonderstellung einnehmen. Während anhand unseres Entscheidungsbaums für alle nicht auf Schwa endenden Ersteinheiten (mit unbetonter letzter Silbe) generell eine Null-Fuge vorausgesagt wird, müssen für auf Schwa auslautende Ersteinheiten sehr komplexe und komplizierte Regelverkettungen modelliert werden.

Am kompliziertesten werden die Voraussagen unseres Modells, wenn Ersteinheiten mit entsprechendem Letztlaut weder über ein Präfix noch über ein Suffix verfügen. Die Prognosen unseres Modells bleiben zwar auch in diesen Fällen sehr zuverlässig, sie beziehen sich aber häufig nur noch auf eine sehr beschränkte Zahl von Erstglied-Lexemen. Dieser Teil des Baums spricht für Erstglieder auf Schwa tendenziell eher gegen eine eindeutige Systematisierbarkeit der Fugenelemente – zumindest gegen eine in linguistische Regeln übertragbare Systematisierbarkeit.

## 5.8 Fazit und Ausblick

Mit den im letzten Kapitel (vgl. 5.7) definierten linguistischen Regeln können die Fugen von etwa 75% der 407 865 Komposita vorausgesagt werden. Für weitere 25% enthält der komplette Entscheidungsbaum ebenfalls Regeln, die es dem Modell möglich machen, insgesamt für 95%[44] der Komposita die Fugen vorauszusagen, allerdings erschließen sich uns bei diesen 25% keine linguistisch sinnvollen Regeln. Das kann einerseits am Unvermögen unsererseits liegen, andererseits resultiert der Verzicht auf die Formulierung linguistischer Regeln in vielen Fällen daraus, dass ein komplexes Zusammenspiel vieler Eigenschaften

[44] Vgl. zu den Details Kapitel 5.6.

im Modell zu Gruppen von Komposita führt, die nur wenige Exemplare enthalten (auch zusätzlich mit nur wenigen unterschiedlichen Erstgliedern) und somit „Regeln" für letztlich lexikalisch definierte Kompositagruppen wirksam zu sein scheinen, die nicht weiter linguistisch abstrahierbar sind.

Demnach stellt sich die Frage: „Gibt es (womöglich nur schwach ausgeprägte) Regularitäten und Systematiken bei der Bildung der Kompositionsfuge" (Michel 2009: 336)? Im Folgenden soll anhand des dargestellten Modells ein Versuch zur Beantwortung dieser Frage (für nominale Komposita) unternommen werden.

Da unser Entscheidungsbaum das Auftreten der jeweiligen Kompositionsfuge mit einer erstaunlich hohen Wahrscheinlichkeit von 95% voraussagt – und die von uns davon abgeleiteten linguistisch sinnvollen Regeln etwa 75% abdecken –, kann die Verteilung der Fugenelemente nicht völlig willkürlich sein, sondern muss auf einer gewissen Systematik basieren. Andernfalls wären derart genaue Voraussagen anhand eines Entscheidungsbaums des geschilderten Typs nicht möglich.

Auch wenn eine gewisse Systematisierbarkeit des Phänomens somit eindeutig bejaht wird, sagt dies noch nicht zwingend etwas darüber aus, ob die Voraussagen unseres Modells aufgrund der Wirksamkeit von Regeln möglich sind, oder ob die Fugenelementverteilung nicht doch zum Teil auf Konventionen – wenn auch gut systematisierbaren Konventionen – basiert. In Verbindung mit den anhand unseres Modells formulierbaren Regularitäten muss zudem prinzipiell berücksichtigt werden, dass in den Entscheidungsbaum nur solche Kriterien eingeflossen sind, die wir messen konnten, und der Baum dementsprechend anders aussehen könnte, wenn weitere – oder einfach andere – Kriterien gemessen würden. Darüber hinaus darf im Rahmen der Beschäftigung mit der Regeln-oder-Konventionen-Frage nicht unerwähnt bleiben, dass die erfolgte Berücksichtigung von Merkmalen wie „Suffix" oder „Letztlaut" in unserem Modell möglicherweise zu einer Überlagerung anderer zugrundeliegender Regeln geführt haben könnte – schließlich lässt sich eine Fuge mit „Wortmaterial" am besten voraussagen, da somit das tatsächliche Lexem bestimmt wird.

Insgesamt kann aufgrund unserer Modellierung, die relativ zu den gemessenen Kriterien ist, weder für die völlige Regelhaftigkeit noch für die ausschließliche Gebundenheit der Fugenelemente an sprachliche Konventionen plädiert werden. Auf der einen Seite spricht die Tatsache, dass bestimmte Gruppen von Erstgliedern, die bei der Kompositabildung mit demselben Fugenelement

auftreten, durch abstrakte gemeinsame Eigenschaften miteinander verbunden sind, für die Wirksamkeit von Regeln – durch die Tatsache, dass die auf der Grundlage des Entscheidungsbaums formulierbaren linguistischen Regeln ca. 75% der ca. 400 000 berücksichtigten Komposita abdecken, wird dieser Eindruck noch verstärkt. Auf der anderen Seite enthält unser Entscheidungsbaum aber auch Fälle, in denen für eine Gruppe von Erstgliedern zwar ein identisches Fugenelement vorhergesagt wird, die Erstglieder selbst aber nicht über gemeinsame abstrakte Eigenschaften verfügen. In solchen Fällen ist keine eindeutige linguistische Regel formulierbar, sondern es können – wenn überhaupt – allenfalls gewisse Tendenzen ausgemacht werden. Als Beispiel sei auf die Voraussage der Null-Fuge für konsonantisch auslautende, suffigierte Erstglieder verwiesen. Auch wenn für eine Teilmenge, genauer gesagt für ca. ein Drittel der entsprechenden Erstglieder, eine abstrakte Regel formuliert werden konnte (Null-Fuge für adjektivische und adverbiale suffigierte Erstglieder), sind für die Gesamtgruppe keine gemeinsamen abstrakten Eigenschaften feststellbar. Da keine linguistisch sinnvolle Regel gefunden wurde, die Erstgliedlexeme wie z.B. *sozial*, *Anerkenntnis* und *Agrarier* gleichermaßen erfasst, ist an dieser Stelle (zumindest teilweise) eher von einer konventionell begründeten Wahl des Fugenelements auszugehen. Die Formulierung einer allgemeinen Regel wird hier, und auch an anderen Stellen im Modell, noch dadurch erschwert, dass die entsprechende Erstglied-Gruppe nicht nur Suffixe im morphologisch engen Sinne enthält, sondern zum Teil auch über einfache „Endungen" (die noch weniger abstrakt beschrieben werden können) definiert ist.

Die zum Teil festgestellte „Regelbasiertheit" der Fugenelemente soll außerdem nicht den Eindruck erwecken, dass innerhalb unseres Modells keine Ausnahmen existieren. Auch wenn der Fokus im Vorhergegangenen eher auf der Beschreibung unserer Fugenmodellierung als auf der Thematisierung von Abweichungen lag, muss erwähnt werden, dass es innerhalb des Entscheidungsbaums keine Regel gibt, zu der keine abweichenden Beispiele generiert wurden. Beispielsweise finden sich im Entscheidungsbaum Bildungen wie *Firmengründer* oder *Themenkomplex*, obwohl für vokalisch auslautende, nicht auf Schwa endende Erstglieder beobachtet werden konnte, dass diese in der Regel ohne Fugenelement auftreten. Als weiterer Abweichungstyp sind zudem Komposita mit variierender Fuge zu nennen, z.B. Bildungen mit dem Erstglied *Unternehmer* (*Unternehmersohn* vs. *Unternehmerssohn*), das aufgrund seiner *er*-Endung gemäß unseres Modells eigentlich ohne Fuge in ein Kompositum integriert werden sollte.

Da die Abweichungen innerhalb unseres Modells jedoch nur im Randbereich (5%) liegen und für die Modellierung allgemein daher nicht problematisch sind, kommen wir dennoch zu dem Ergebnis, dass sich „– bei allen noch bestehenden Schwankungen – einige generelle Regelungen fixieren [lassen], mit denen die Wahlmöglichkeiten stark eingeschränkt werden, wenn auch nicht in jedem Fall eine eindeutige Voraussagbarkeit der Fugengestaltung gegeben ist" (Fleischer/Barz 1995: 137f.).

Abgesehen von der Konventionen-versus-Regeln-Frage können anhand unseres Entscheidungsbaums auch Aussagen über die Relevanz einzelner linguistischer Merkmale für das Auftreten von Fugenelementen getroffen werden. Dabei muss zwischen konsonantisch und vokalisch auslautenden Erstgliedern unterschieden werden. Wie bereits erwähnt, sind die Fugenvorhersagen für konsonantisch auslautende Erstglieder deutlich weniger komplex als die für vokalisch auslautende Erstglieder: Es wurde gezeigt, dass bei Ersteinheiten mit einem Konsonanten als Letztlaut allein aufgrund der „Endung" relativ verlässliche, unmittelbare Aussagen über die Gestaltung der Kompositionsfuge getroffen werden können. Darüber hinaus ist festzustellen, dass lautliche Aspekte hier eine recht dominante Rolle spielen. Für Ersteinheiten mit einem Vokal als Letztlaut gestalten sich die Prognosen insofern generell schwieriger, als die Fugenvorhersagen nur mit Hilfe von relativ komplexen Verkettungen von Regeln getroffen werden können. Abgesehen von lautlichen Aspekten, die auch hier eine wichtige Rolle spielen, stellt das Flexionsparadigma des Erstglieds ein sehr „mächtiges" Merkmal dar. Für beide Erstgliedtypen zeigt unser Modell deutlich, dass die Wahl des Fugenelements – zumindest für die hier betrachteten, nominalen Komposita – in so gut wie keinem Zusammenhang mit Eigenschaften des Zweitglieds steht, sondern – um mit Lohde (2006: 22) zu sprechen – „grundsätzlich vom Charakter des Erstgliedes der Komposition" abhängig ist.

Die Erforschung der Fugen in Komposita mit Methoden des maschinellen Lernens steht erst am Anfang. Vor allem was die Auswahl der gemessenen Eigenschaften betrifft, sind weitere Analysen notwendig: Basierend auf den Hypothesen der bisherigen Forschung zu den Fugenelementen ließen sich noch weitere Eigenschaften integrieren, z.B. aus dem Bereich der Semantik. Zudem könnten Eigenschaften, die in unserer Studie nur indirekt gemessen wurden – wie z.B. die Frage, ob eine bestimmte Fuge für ein Erstglied paradigmisch ist – direkt in die Modellierung einbezogen werden, was die Interpretation des Entscheidungsbaumes vereinfachen würde.

In der vorliegenden Studie beschränkten wir uns zudem auf Komposita, die bezüglich Fuge nicht (oder kaum) variieren. In einer Folgestudie soll aber dieser Aspekt der Variation genauer untersucht werden: Welche Faktoren sind es, die zu Variation führen und mit welchen Faktoren lässt sich eine bestimmte Variation voraussagen?

Wir sehen die Chance des maschinellen Lernens darin, eine sehr große Menge von Phänomenen systematisch untersuchen und so zu grammatischen Regeln kommen zu können, die eine breite empirische Basis beschreiben.[45] Dabei besteht die Möglichkeit, sowohl bestehende Hypothesen zu überprüfen als auch durch die komplexe Kombination mehrerer Eigenschaften der Phänomene auf neue Regeln zu stoßen.

[45] Zu relativieren ist die hier konstatierte Chance, die die Anwendung maschineller Lernverfahren zweifelsohne bietet, insofern, als sowohl die Festlegung der zu messenden Kriterien als auch die linguistische Auswertung der automatisch generierten Entscheidungsbäume in der praktischen Umsetzung mit einem hohen Zeitaufwand verbunden sind.

## Literatur


Albert, Ruth/Koster, Cor J. (2002): Empirie in Linguistik und Sprachlehrforschung. Ein methodologisches Arbeitsbuch. Tübingen: Narr.

Althaus, Hans Peter/Henne, Helmut/Wiegand, Herbert Ernst (Hg.) (1980): Lexikon der Germanistischen Linguistik. 2. vollst. neu bearb. u. erw. Aufl. Tübingen: Niemeyer.

Altmann, Hans (2011): Prüfungswissen Wortbildung. 3. Aufl. (= UTB Sprach-/Literaturwissenschaft 3458). Göttingen u.a.: Vandenhoeck & Ruprecht.

Ammon, Ulrich (1995): Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten. Berlin/New York: de Gruyter.

Ammon, Ulrich (2005): Standard und Variation: Norm, Autorität, Legitimation. In: Eichinger/Kallmeyer (Hg.), 28-40.

Auer, Peter (2000): *On line*-Syntax – Oder: was es bedeuten könnte, die Zeitlichkeit der mündlichen Sprache ernst zu nehmen. In: Sprache und Literatur 85: 43-56.

Augst, Gerhard (1975): Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache. (= Forschungsberichte des Instituts für deutsche Sprache 25). Tübingen: Narr.

Baayen, R. Harald/Piepenbrock, Richard/Gulikers, L. (1995): The CELEX Lexical Database (CD-ROM). Philadelphia.

Bank of English (BoE): http://www.mycobuild.com/about-collins-corpus.aspx (Stand: September 2011).

Barbour, Stephen (2005): Standardvariation im Deutschen und im Englischen: Auswirkungen auf die Kommunikation zwischen Sprechern beider Sprachen. In: Eichinger/Kallmeyer (Hg.), 324-333.

Barbour, Stephen/Stevenson, Patrick (1998): Variation im Deutschen. Soziolinguistische Perspektiven. Berlin, New York: de Gruyter.

Belica, Cyril/Kupietz, Marc/Witt, Andreas/Lungen, Harald (2009): The morphosyntactic annotation of DeReKo: interpretation, opportunities, and pitfalls – IV Einblicke in die aktuelle Forschung/Insights into current studies. In: Konopka et al. (Hg.), 451-469.

Berruto, Gaetano (2010): Identifying dimensions of linguistic variation in a language space. In: Auer, Peter/Schmidt, Jürgen Erich (Hg.): Language and space. An international handbook of linguistic variation. Bd. 1: Theories and methods. (= HSK 30.1). Berlin/New York: de Gruyter, 226-240.

Biber, Douglas (1993): Representativeness in corpus design. In: Literary and Linguistic Computing 8, 4: 243-257.

Biber, Douglas (2010): What can a corpus tell us about registers and genres? In: O'Keeffe/McCarthy (Hg.), 241-244.

Biber, Douglas/Conrad, Susan (2009): Register, genre and style. Cambridge: University Press.

Biber, Douglas/Johansson, Stig/Leech, Geoffrey/Conrad, Susan/Finegan, Edward (1999): Longman Grammar of Spoken and Written English. Harlow: Longman.

Biber, Douglas/Johansson, Stig/Leech, Geoffrey/Conrad, Susan/Finegan, Edward (2006): Longman Grammar of Spoken and Written English. 5. Aufl. Harlow: Longman.

Bird, Steven/Chen, Yi/Davidson, Susan B./Lee, Haejoong/Zhen, Yifeng (2005): Extending XPath to support linguistic queries. Proceedings of the Workshop on Programming Language Technologies for XML (Plan-X). Philadelphia: University of Pennsylvania/Melbourne: University of Melbourne. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1136&context=cis_papers (Stand: August 2013).

British National Corpus (BNC XML): http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html (Stand: August 2011).

Burnage, Gavin (1990): CELEX: A Guide for users. Appendix 2: "Computer phonetic character codes". Nijmegen: Centre for Lexical Information.

Bußmann, Hadumod (Hg.) (2008): Lexikon der Sprachwissenschaft. 4., durchges. u. bibliogr. erg. Aufl. Stuttgart: Kröner.

Carstensen, Kai-Uwe/Ebert, Christian/Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf (Hg.) (2001): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg/Berlin: Spektrum Akademischer Verlag.

Chen, Peter Pin-Shan (1976): The entity-relationship model – toward a unified view of data. In: ACM Transactions on Database Systems 1/1/1976, ACM-Press, 9-36.

Chiarcos, Christian/Dipper, Stefanie/Götze, Michael/Leser, Ulf/Lüdeling, Anke/Ritz, Julia/Stede, Manfred (2008): A flexible framework for integrating annotations from different tools and tag sets. In: Traitement Automatique des Langues 49, 2: 271-293.

Christiansen, Tom/Foy, Brian D./Wall, Larry/Orwant, Jon (2000): Programming Perl. Sebastopol: O'Reilly.

Church, Kenneth W./Gale, William A. (1995): Poisson mixtures. In: Natural Language Engineering 1, 2: 163-190.

Church, Kenneth W./Mercer, Robert L. (1993): Introduction to the special issue on computational linguistics using large corpora. In: Computational Linguistics 19, 1: 1-24.

Clancy, Brian (2010): Building a corpus to represent a variety of a language. In: O'Keeffe/McCarthy (Hg.), 80-92.

Conrad, Susan (2010): What can a corpus tell us about grammar? In: O'Keeffe/McCarthy (Hg.), 227-240.

Davies, Mark (2005): The advantage of using relational databases for large corpora. Speed, advanced queries, and unlimited annotation. In: International Journal of Corpus Linguistics 10, 3: 307-334.

Dean, Jeffrey/Ghemawat, Sanjay (2004): MapReduce: simplified data processing on large clusters. In: Proceedings of OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December 2004. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/de//archive/mapreduce-osdi04.pdf (Stand: Januar 2013).

Di Meola, Claudio (2009): Rektionsschwankungen bei Präpositionen – erlaubt, verboten, unbeachtet. In: Konopka/Strecker (Hg.), 195-221.

Donalies, Elke (2011): Grammatische Variation im Deutschen: *Tagtraum, Tageslicht, Tagedieb.* Ein korpuslinguistisches Experiment zu variierenden Wortformen und Fugenelementen in zusammengesetzten Substantiven. Mit einem Exkurs und zahlreichen Statistiken von Noah Bubenhofer. (= amades – Arbeitspapiere und Materialien zur deutschen Sprache 42) Mannheim: Institut für Deutsche Sprache.

Dovalil, Vit (2006): Sprachnormenwandel im geschriebenen Deutsch an der Schwelle zum 21. Jahrhundert. Frankfurt a.M. u.a.: Peter Lang.

Duden (1999): Duden – Das Große Wörterbuch der Deutschen Sprache in zehn Bänden. Hrsg. vom Wissenschaftlichen Rat der Dudenredaktion. 3. vollst. neubearb. Aufl. (= Duden 8). Mannheim: Bibliographisches Institut.

Duden (2005): Duden – Die Grammatik. 7. völlig neu erarb. und erw. Aufl. (= Duden 4). Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Duden (2009): Duden – Die Grammatik: unentbehrlich für richtiges Deutsch. 8. überarb. Aufl. (= Duden 4). Mannheim/Wien/Zürich: Dudenverlag.

Dürscheid, Christa/Elspaß, Stephan/Ziegler, Arne (2011): Grammatische Variabilität im Gebrauchsstandard: das Projekt „Variantengrammatik des Standarddeutschen". In: Konopka et al. (Hg.), 123-140.

Durell, Martin (1999): Standardsprache in England und Deutschland. In: Zeitschrift für germanistische Linguistik 27: 285-308.

DWDS: http://www.dwds.de/resource/kerncorpus/ (Stand: August 2011).

Eichinger, Ludwig M. (2005a): Standardnorm, Sprachkultur und die Veränderung der normativen Erwartungen. In: Eichinger/Kallmeyer (Hg.), 363-381.

Eichinger, Ludwig M. (2005b): Norm und regionale Variation. Zur realen Existenz nationaler Varietäten. In: Lenz/Mattheier (Hg.), 141-162.

Eichinger, Ludwig M./Kallmeyer, Werner (Hg.) (2005): Standardvariation. Wie viel Variation verträgt die deutsche Sprache? Jahrbuch 2004 des Instituts für deutsche Sprache. Berlin/New York: de Gruyter.

Eisenberg, Peter (2007): Sprachliches Wissen im Wörterbuch der Zweifelsfälle. Über die Rekonstruktion einer Gebrauchsnorm. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur 3: 209-228.

Elspaß, Stephan (2005a): Zum sprachpolitischen Umgang mit regionaler Variation der Standardsprache. In: Kilian, Jörg (Hg.): Sprache und Politik. Deutsch im demokratischen Staat (= Thema Deutsch 6). Mannheim/Leipzig/Wien/Zürich: Dudenverlag, 294-313.

Elspaß, Stephan (2005b): Standardisierung des Deutschen. Ansichten aus der neueren Sprachgeschichte 'von unten'. In: Eichinger/Kallmeyer (Hg.), 63-99.

Evert, Stefan (2006): How random is a corpus? The library metaphor. In: Zeitschrift für Anglistik und Amerikanistik 54, 2: 177-190.

Fanselow, Gisbert/Fery, Caroline/Schlesewsky, Matthias/Vogel, Ralf (Hg.) (2006): Gradience in grammar. Generative perspectives. Oxford: University Press.

Fiehler, Reinhard (2011): Korpusbasierte Analyse von Univerbierungsprozessen. In: Konopka et al. (Hg.), 141-155.

Fleischer, Wolfgang (1971): Wortbildung der deutschen Gegenwartssprache. 2. unveränd. Aufl. Tübingen: Niemeyer.

Fleischer, Wolfgang/Barz, Irmhild (1995): Wortbildung der deutschen Gegenwartssprache. 2., durchges. und erg. Aufl. Tübingen: Niemeyer.

Fuhrhop, Nanna (2000): Zeigen Fugenelemente die Morphologisierung von Komposita an? In: Fuhrhop, Nanna/Thieroff, Rolf/Teuber, Oliver/Tamrat, Matthias (Hg.): Deutsche Grammatik in Theorie und Praxis: Aus Anlaß des 60. Geburtstags von Peter Eisenberg am 18. Mai 2000. Tübingen: Niemeyer, 201-214.

Fuhrhop, Nanna (1996): Fugenelemente. In: Lang, Ewald/Zifonun, Gisela (Hg.): Deutsch – typologisch. Jahrbuch 1995 des Instituts für Deutsche Sprache. Berlin/New York: de Gruyter, 525-550.

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (Hg.): Collocations and idioms: linguistic, lexicographic, and computational aspects. London: Continuum Press, 23-41.

Glinz, Hans (1980): Deutsche Standardsprache der Gegenwart. In: Althaus/Henne/Wiegand (Hg.), 609-619.

*grammis*: Grammatisches Informationssystem. Mannheim: Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/index.html (Stand: Januar 2013).

Gries, Stefan Thomas (2008a): Dispersions and adjusted frequencies in corpora. In: International Journal of Corpus Linguistics 13, 4: 403-437.

Gries, Stefan Thomas (2008b): Statistik für Sprachwissenschaftler. (= Studienbücher zur Linguistik 13). Göttingen: Vandenhoeck & Ruprecht.

Gries, Stefan Thomas (2009): Dispersions and adjusted frequencies in corpora: further explorations. In: Language and Computers 71, 1: 197-212.

Haugen, Einar (1994): Standardization. In: Asher, Ronald E. (Hg.): The encyclopedia of language and linguistics. 10-volume set. Bd. 8. Oxford/New York: Pergamon Press, 4340-4342.

Hennig, Mathilde/Müller, Christoph (Hg.) (2009): Wie normal ist die Norm? Sprachliche Normen im Spannungsfeld von Sprachwissenschaft, Sprachöffentlichkeit und Sprachdidaktik. Kassel: University Press.

Heringer, Hans Jürgen (2011): *Anfang diesen Jahres?* — Anders gefragt. In: *grammis*: Grammatik in Fragen und Antworten. Mannheim: Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/call/public/fragen.ansicht?v_typ=e&v_id=4524 (Stand: Januar 2013).

Hundt, Marianne (2008): Text corpora. In: Lüdeling/Kytö (Hg.), 168-186.

Hunston, Susan (2002): Corpora in applied linguistics. Cambridge: University Press.

Hunston, Susan (2008): Collection strategies and design decisions. In: Lüdeling/Kytö (Hg.), 154-167.

Ihaka, Ross/Gentleman, Robert (1996): R: a language for data analysis and graphics. In: Journal of Computational and Graphical Statistics 5, 3: 299-314.

Institut für Deutsche Sprache (2011): Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2011-I (Release vom 29.03.2010). http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html (Stand: Januar 2013).

International Corpus of English (ICE): http://ice-corpora.net/ice/design.htm (Stand: August 2011).

Jäger, Siegfried (1980): Standardsprache. In: Althaus/Henne/Wiegand (Hg.), 375-379.

John, George H. (1997): Enhancements to the data mining process. Univ. Diss., Dept. of Computer Science, Stanford University.

Kepser, Stephan/Mönnich, Uwe/Morawietz, Frank (2010): Regular query techniques for XML-documents. In: Witt, Andreas/Metzing, Dieter (Hg.): Linguistic modeling of information and markup languages. Contributions to language technology. Heidelberg/London/New York: Springer, 249-266.

Kilgarriff, Adam (2001): Comparing corpora. In: International Journal of Corpus Linguistics 6, 1: 1-37.

Klein, Wolfgang (2004): Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS). In: Scharnhorst, Jürgen (Hg.). Sprachkultur und Lexikographie: von der Forschung zur Nutzung von Wörterbüchern. Frankfurt a.M./Berlin/Wien: Peter Lang, 281-311.

Koch, Peter/Oesterreicher, Wulf (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: Romanistisches Jahrbuch 36: 15-43.

Koch, Peter/Oesterreicher, Wulf (2008): Mündlichkeit und Schriftlichkeit von Texten. In: Janich, Nina (Hg.): Textlinguistik. 15 Einführungen. Tübingen: Narr, 199-215.

König, Werner (2004): dtv-Atlas Deutsche Sprache. Mit 155 Abbildungskarten in Farbe. 14., durchges. u. aktual. Aufl. München: dtv.

Konopka, Marek (2011): Grammatik verstehen lernen und korpusgestützte Analysen von Zweifelsfällen. In: Köpcke, Klaus-Michael/Ziegler, Arne (Hg.): Grammatik – Lehren, Lernen, Verstehen. Zugänge zur Grammatik des Gegenwartsdeutschen. (= Reihe Germanistische Linguistik 293). Berlin/New York: de Gruyter, 265-285.

Konopka, Marek (2010): Niedrigfrequente grammatische Phänomene als sprachliche Zweifelsfälle. In: Korpus – grammatika – axiologie 2: 24-44.

Konopka, Marek/Strecker, Bruno (Hg.) (2009): Deutsche Grammatik – Regeln, Normen, Sprachgebrauch. Jahrbuch 2008 des Instituts für Deutsche Sprache. Berlin/New York: de Gruyter.

Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Štícha, František/Waßner, Ulrich H. (Hg.) (2011): Grammatik und Korpora 2009. Dritte internationale Konferenz, Mannheim, 22.-24.09.2009 (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen: Narr.

Kubczak, Jacqueline (2008): *Dank seines Einflusses* oder *dank seinem Einfluss*, *dank deiner* oder *dank dir*? – Kasus nach *dank*. In: *grammis*: Grammatik in Fragen und Antworten. Mannheim: Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/call/public/fragen.ansicht?v_id= 69 (Stand: Januar 2013).

Kubczak, Jacqueline/Konopka, Marek (2008): Grammatical variation in Near-Standard German: a corpus-based project at the Institute for the German Language (IDS) in Mannheim. In: Šticha, František/Fried, Mirjam (Hg.): Grammar & Corpora 2007. Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice. Praha: Academia, 251-260.

Kullback, Solomon/Leibler, Richard A. (1951): On Information and sufficiency. In: The Annals of Mathematical Statistics 22, 1: 79-86.

Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawa-

guchi, Yuji (Hg.): Working papers in corpus-based linguistics and language education 3: 53-59.

Kupietz, Marc/Belica, Cyril/Keibel, Holger (2010): The German Reference Corpus DeReKo: a primordial sample for linguistic research. In: Proceedings of the seventh conference on International Language Resources and Evaluation. Valletta, Malta, 1848-1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (Stand: Januar 2013).

Leech, Geoffrey N./Rayson, Paul/Wilson, Andrew (2001): Word frequencies in written and spoken English: based on the British National Corpus. Harlow/München: Longman.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen: Narr.

Lenz, Alexandra N./Mattheier, Klaus J. (Hg.) (2005): Varietäten – Theorie und Empirie. Frankfurt a.M.: Peter Lang.

Lijffijt, Jefrey/Gries, Stefan Thomas (2012): Correction to "Dispersions and adjusted frequencies in corpora". In: International Journal of Corpus Linguistics 16, 1: 147-149.

Lin, Jimmy/Dyer, Chris (2010): Data-intensive text processing with MapReduce. Synthesis lectures on human language technologies. San Rafael, CA: Morgan & Claypool Publishers.

Löffler, Heinrich (2005): Wieviel Variation verträgt die deutsche Standardsprache? Begriffsklärung: Standard und Gegenbegriffe. In: Eichinger/Kallmeyer (Hg.), 7-27.

Lohde, Michael (2006): Wortbildung des modernen Deutschen: ein Lehr- und Übungsbuch. Tübingen: Narr.

Lüdeling, Anke/Kytö, Merja E. (Hg.) (2008): Corpus linguistics. An international handbook. Bd. 1: Theories and methods. (= Handbücher zur Sprach- und Kommunikationswissenschaft 29.1). Berlin/New York: de Gruyter.

Lüdtke, Jens/Mattheier, Klaus J. (2005): Variation – Varietäten – Standardsprachen. Wege für die Forschung. In: Lenz/Mattheier (Hg.), 13-38.

Majumder, Anirban/Rastogi, Rajeev/Vanama, Sriram (2008): Scalable regular expression matching on data streams. Proceedings of the 2008 ACM SIGMOD international conference on management of data. Vancouver: ACM, 161-172.

Manning, Christopher D./Schütze, Hinrich (2002): Foundations of statistical natural language processing. 5. Aufl. Cambridge: MIT Press.

Michel, Sascha (2009): *Schaden-0-ersatz* vs. *Schaden-s-ersatz*. Ein Erklärungsansatz synchroner Schwankungsfälle bei der Fugenbildung von N+N-Komposita. In: Deutsche Sprache 4: 334-351.

Nelson, Mike (2010): Building a written corpus: what are the basics? In: O'Keeffe/McCarthy (Hg.), 53-65.

Nübling, Damaris/Szczepaniak, Renata (2011): *Merkmal(s?)analyse*, *Seminar(s?)arbeit* und *Essen(s?)ausgabe*: Zweifelsfälle der Verfugung als Indikatoren für Sprachwandel. In: Zeitschrift für Sprachwissenschaft 30: 45-73.

Oakes, Michael (1998): Statistics for corpus linguistics. (= Edinburgh Textbooks in Empirical Linguistics). Edinburgh: Edinburgh University Press.

Ortner, Lorelies/Müller-Bollhagen, Elgin/Ortner, Hanspeter (1991): Fugen: Die formale Verbindung zwischen den Konstituenten (mit und ohne Fugenelement). In: Deutsche Wortbildung, Typen und Tendenzen in der Gegenwartssprache, Eine Bestandsaufnahme des Instituts für deutsche Sprache, Forschungsstelle Innsbruck. Vierter Hauptteil: Substantivkomposita. (= Sprache der Gegenwart 79). Berlin/New York: de Gruyter, 50-111.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. (= UTB 3433). Paderborn: Fink.

Pomikálek, Jan/Rychlý, Pavel/Kilgarriff, Adam (2009): Scaling to billion-plus word corpora. In: Advances in Computational Linguistics 41: 3-13.

Quinlan, J. Ross (1993): C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann.

Ramers, Karl Heinz (1997): Die Kunst der Fuge: Zum morphologischen Status von Verbindungselementen in Nominalkomposita. In: Dürscheid, Christa/Ramers, Karl-Heinz/Schwarz, Monika (Hg.): Sprache im Fokus: Festschrift für Heinz Vater zum 65. Geburtstag. Tübingen: Niemeyer, 33-46.

Rehm, Georg/Schonefeld, Oliver/Witt, Andreas/Chiarcos, Christian/Lehmberg, Timm (2008): A web-platform for preserving, exploring, visualising and querying linguistic corpora and other resources. In: Procesamiento del Lenguaje Natural 41: 155-162.

Scherer, Carmen (2006): Korpuslinguistik. Heidelberg: Winter.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf (Stand: Januar 2013).

Schneider, Roman (2009): Information Retrieval mit Oracle Text. In: iX – Magazin für professionelle Informationstechnik 9: 144-147.

Schneider, Roman (2012): Evaluating DBMS-based access strategies to very large multi-layer annotated corpora. In: Proceedings of the LREC-2012 Workshop "Challenges in the management of large corpora", 22.05.2012, Istanbul.

Shearer, Colin (2000): The CRISP-DM model: the new blueprint for data mining. In: Journal of Data Warehousing 5, 4: 13-22.

Sheskin, David J. (2007): Handbook of parametric and nonparametric statistical procedures. 4. Aufl. Boca Raton/London/New York: Chapman & Hall/CRC.

Sinclair, John (2004): Corpus and text – basic principles. In: Wynne, Martin (Hg.): Developing linguistic corpora: a guide to good practice. Oxford: ADHS. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm (Stand: Januar 2013).

Stonebraker, Michael/Abadi, Daniel/Dewitt, David J./Madden, Sam/Paulson, Erik/Pavlo, Andrew/Rasin, Alexander (2010): MapReduce and parallel DBMSs: friends or foes? In: Communications of the ACM, January 2010, 53, 1: 64-71.

Strecker, Bruno (2010a): *Anfang diesen Jahres* oder *Ende dieses Jahres*? – Genitiv Singular beim Demonstrativ-Artikel. In: *grammis*: Grammatik in Fragen und Antworten. Mannheim: Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/call/public/fragen.ansicht?v_id=36 (Stand: Januar 2013).

Strecker, Bruno (2010b): Wörter gibt's, die gibt's gar nicht! Ein Exkurs ins grammatische Raritätenkabinett. In: Korpus – Grammatika – Axiologie 2: 57-64.

Strecker, Bruno (2011): Korpusgrammatik zwischen reiner Statistik und „intelligenter" Grammatikografie. In: Konopka et al. (Hg.), 23-45.

Weiß, Christian (2005): Die thematische Erschließung von Sprachkorpora. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2005). Mannheim: Institut für Deutsche Sprache.

Wellmann, Hans/Reindl, Nikolaus/Fahrmeier, Annemarie (1974): Zur morphologischen Regelung der Substantivkomposition im heutigen Deutsch. In: Zeitschrift für deutsche Philologie 93: 358-378.

Wiese, Bernd (2009): Variation in der Flexionsmorphologie: Starke und schwache Adjektivflexion nach Pronominaladjektiven. In: Konopka/Strecker (Hg.), 166-194.

Wiesinger, Peter (1983): Die Einteilung der deutschen Dialekte. In: Besch, Werner/Knoop, Ulrich/Putschke, Wolfgang/Wiegand, Herbert Ernst (Hg.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Bd. 2. (= Handbücher zur Sprach- und Kommunikationsforschung 1.2). Berlin/New York: de Gruyter, 807-900.

Witten, Ian H./Frank, Eibe/Hall, Mark A. (2005): Data mining: practical machine learning tools and techniques. 2. Aufl. San Francisco: Morgan Kaufmann.

Xiao, Richard (2008): Well-known and influential corpora. In: Lüdeling/Kytö (Hg.), 383-457.

Zeldes, Amir/Ritz, Julia/Lüdeling, Anke/Chiarcos, Christian (2009): ANNIS: a search tool for multi-layer annotated corpora. In: Proceedings of Corpus Linguistics 2009. July 20-23, Liverpool, UK. [Manuskript]. https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/amir/pdf/CL2009_ANNIS_pre.pdf (Stand: August 2013).

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno/Ballweg, Joachim et al. (1997): Grammatik der deutschen Sprache. (= Schriften des Instituts für Deutsche Sprache 7.1-7.3). Berlin: de Gruyter.

# Anhang

## Beispielaussagen zum Thema Standardsprache

Ammon (2005):

> Aber auch in topischer, also regionaler Hinsicht bedeutet Standardsprachlichkeit keine völlige Einheitlichkeit – ob auch nicht in stratischer, also sozialer Hinsicht, möchte ich offen lassen. [...] Im Moment setze ich voraus, dass sie als Leser die Tatsache einsehen, dass Standarddeutsch regional variiert [...].
>
> Die regionale Einheitsidee ist allerdings insofern richtig, als auf standardsprachlicher Ebene die Regionen, zwischen denen variiert wird, durchschnittlich größer sind als im Dialekt. Außerdem ist die Zahl variierender sprachlicher Einheiten weitaus beschränkter. Dies entspricht einem der Hauptzwecke sprachlicher Standardisierung, nämlich; dialektale Kommunikationsschranken zu überbrücken. (29)
>
> Der *Standard* wird auch oft mit der *Norm* der Sprache schlechthin gleichgesetzt. Die Aussage „Das ist normwidrig im Deutschen" meint dann einfach: 'Es ist kein Standarddeutsch' [...]. (31)
>
> Im Gegensatz zu Nonstandardvarietäten sind Standardvarietäten
>
> a) kodifiziert, d. h. es gibt für sie Sprachkodizes oder -kodexe im Sinne autoritativer Nachschlagewerke für den korrekten Gebrauch. Sie werden
>
> b) förmlich gelehrt, und sie haben
>
> c) amtlichen Status, schon durch die Schule, aber meist darüber hinaus.
>
> Als Folge davon wird
>
> a) die Einhaltung der Normen kontrolliert von Sprachnormautoritäten von Berufs wegen, zu deren Berufsaufgaben die Korrektur von Sprachfehlern gehört, z.B. Lehrer oder Vorgesetzte auf Ämtern.
>
> In diesem Sinne sind Standardvarietäten *förmlich institutionalisierte Vorschriften* – eine allerdings missverständliche Formulierung, für deren konstruktive Kritik ich dankbar wäre. (32)

Barbour (2005):

> Ich gehe insbesondere von einer Auffassung des Begriffes Standardsprache aus, die in der britischen Soziolinguistik verbreitet ist; nach diesem Verständnis des Begriffes ist ein „standard language" der in der Gesellschaft sowohl in geschriebener als auch in gesprochener Form als maßgeblich akzeptierter Sprachgebrauch. Nach diesem Verständnis des Phänomens ist Standardsprache (ich benutze aus Bequemlichkeitsgründen den deutschen Terminus, gemeint ist aber hier durchweg eher der englische Begriff) der gesamte Sprachgebrauch einer

> Bildungsschicht, d. h. sie umfasst alle sprachlichen Register dieser Schicht, und schließt demnach das ein, was man im Deutschen als Umgangssprache der Bildungsschicht bezeichnen würde. [...]. Es handelt sich hier sowohl um gesprochene als auch um geschriebene Sprache, und sowohl um formelle als auch um informelle Sprache, also auch um Umgangssprache. Standardisiert ist diese Varietät vor allem in der Grammatik und der Orthographie, kaum in der Aussprache, und nur bedingt in der Lexik. Da es sowohl um gesprochene als auch um geschriebene Sprache geht, handelt es sich hier um einen Gebrauchsstandard oder eine Gebrauchsnorm, und nicht nur um einen kodifizierten Standard oder eine Sollnorm [...]. (325)

Barbour/Stevenson (1998):

> Die Erforschung der deutschen Sprache in den deutschsprachigen Ländern hat sich in den meisten Fällen als ein Studium von zwei sich deutlich voneinander abgrenzenden **Sprachtypen**[1] erwiesen. [...]. Der erste dieser Sprachtypen wird im Deutschen unter mehreren Bezeichnungen geführt. Man nennt ihn z.B. *Hochsprache,* **Schriftsprache**, *Literatursprache, Einheitssprache* oder auch *Standardsprache.* Es ist jene Art von Deutsch, die man traditionell sowohl mündlich als auch schriftlich an Schulen erwartet und weiterentwickelt, die in den meisten Grammatiken und Wörterbüchern zu finden ist und so gut wie allen ausländischen Deutschlernenden beigebracht wird. Die Aussprache der Standardvarietät wird ebenso wie ihre Grammatik und Orthographie u.a. im *Duden*, und darüber hinaus in *Siebs Deutsche*[r] *Aussprache* beschrieben. Gemessen an seinem sozialen Prestige gilt Standarddeutsch vielfach, doch nicht grundsätzlich, als die akzeptabelste Sprachform. (145)

Bußmann (2008):

> **Standardsprache** [Auch: Hochsprache, Nationalsprache]. Seit den 70er Jahren in Deutschland übliche legitimierte, überregionale, mündliche und schriftliche Sprachform der sozialen Mittel- bzw. Oberschicht; in diesem Sinn synonyme Verwendung mit der (wertenden) Bezeichnung „Hochsprache". Entsprechend ihrer Funktion als öffentliches Verständigungsmittel unterliegt sie (besonders in den Bereichen Grammatik, Aussprache und Rechtschreibung) weitgehender Normierung, die über öffentliche Medien und Institutionen, vor allem aber durch das Bildungssystem kontrolliert und vermittelt wird. (680)

Duden (1999):

> **Standardsprache** [...]: über den Mundarten, lokalen Umgangssprachen u. Gruppensprachen stehende, allgemein verbindliche Sprachform; gesprochene u. geschriebene Erscheinungsform der Hochsprache. (3699)

[1] Fettdruck im Original, Kursivsetzung durch die Herausgeber.

Durell (1999):

> Der Terminus Standardsprache hat sich als wissenschaftssprachliche Lehnübersetzung des englischen standard language seit etwa 1970 auch in der deutschsprachigen soziolinguistischen Literatur eingebürgert zur Bezeichnung der Prestigevarietät. Er ist vor allem nützlich, weil er sich auf ein weiteres Variantenspektrum beziehen läßt als der gemeinsprachliche deutsche Terminus Hochsprache, der oft ausschließlich mit der Sprache der Literatur identifiziert wird [...]. (285)
>
> Die aus einem Selektionsprozeß entstandene Standardsprache muß dann von allen Sprachteilhabern als alleingültige überregionale (bzw. nationale oder offizielle) Varietät in der Sprachgemeinschaft akzeptiert werden. Im typischen modernen Nationalstaat erfolgt in ihr die primäre Alphabetisierung, sie ist die Sprache der Rechtsprechung, der Behörden, und, mit sehr wenigen Ausnahmen, des gesamten Schriftverkehrs. (288)
>
> Entsprechend dem Ursprung der Standardsprachen als Bildungsnormen, als Sprachen einer Bildungselite, was in diesem Fall für Deutsch und Englisch in gleicher Weise gilt, genießen sie hohes Ansehen auch unter denjenigen Bevölkerungsschichten, die sie vielleicht nur unvollkommen (wenn überhaupt) beherrschen [...]. (289)
>
> Die Kodifizierung von Standardsprachen hängt charakteristisch sehr eng mit der Erschaffung eines Nationalstaates zusammen. In dem diese Sprache eine symbolische Funktion als Ausdruck der nationalen Einheit ausübt. (296)
>
> Dagegen kennt das Englische eine volle Elaboriertheit der sprachlichen Funktionen in der Standardsprache, während die deutsche Standardsprache erst in relativ neuerer Zeit auf den alltäglichen mündlichen Gebrauch ausgedehnt wurde und das informelle Register noch nicht universell als echt standardsprachlich akzeptiert ist. (304)

Eichinger (2005a):

> Die Norm der deutschen Standardsprache existiert offenkundig, einigermaßen wissen wir auch über ihre Grenzen Bescheid. Nicht zuletzt im Verlauf der in diesem Band dokumentierten Tagung ist aber klar geworden, dass Variation, ja die Möglichkeit zur Variation zur Standardsprachlichkeit gehören. (363)
>
> Als Vorbilder eines standardsprachlichen Schreibens und Sprechens werden allmählich die Sprachen der jeweiligen Leitmedien angesehen und akzeptiert. Das ist zunächst die Sprache, wie sie sich in den Zeitungen findet, später ist es das, was der Rundfunk und das Fernsehen an sprachlichen Formen mit sich bringen. (368)
>
> Zum einen meint Standardsprachlichkeit, dass man den Festschreibungen auf den verschiedenen Ebenen der sprachlich-formalen Beschreibung von Aussprache bis Flexion folgt, in diesem Sinn keine Fehler macht. Auf einer zweiten

> Ebene geht es dann allerdings schon darum, dass die Auswahl aus den Optionen, die sich hier im Prinzip bieten diamedial, aussageintentional und textsortenspezifisch variieren. Und zum dritten ist offenkundig, dass mit der Prätention auf standardsprachliches Verhalten ein sozialsymbolischer Anspruch verbunden ist, der ebenfalls die Beurteilung bestimmter Äußerungsformen steuert. (379f.)

Eichinger (2005b):

> Es ist wenig kontrovers, von den europäischen Kultursprachen, also dem Deutschen, Italienischen, Französischen oder Spanischen und weiteren zu sagen, sie hätten den Status der Standardsprachlichkeit erreicht, also jenen Zustand, in dem, um es direkt auf das deutsche Beispiel hin zu formulieren, eine als Druck- und Schriftsprache entwickelte Form auch weite Räume der öffentlichen und halböffentlichen Mündlichkeit eingenommen und dadurch die Sprechsprachen der Vergangenheit, die „Mundarten", marginalisiert hat. (141f.)
>
> Es ist aber offenkundig, dass mit den medialen Umbrüchen in der zweiten Hälfte des 20. Jahrhunderts die tradierte schriftsprachliche Norm weniger und weniger als die alleinige Basis sozial angemessenen und sprachlichen Verhaltens angesehen, sondern dass umgekehrt Schriftlichkeit nunmehr auch auf verschiedene mündliche Praxen rückbezogen wird – gerade das Fernsehen kennt eine Menge von so gelagerten Formen. Dadurch verliert die schriftliche Standardnorm ihre strikte Dominanz. Das wirkt aus dem Grund nicht so auffällig, da diese Norm bei weitem nicht in allen Bereichen voll ausformuliert ist – und das betrifft nicht nur Wortschatzfragen. Damit gibt es einen erheblichen Spielraum, in dem die Ausgestaltung des Standards im Einzelnen der Übereinkunft der normgebenden Schichten und Instanzen entspricht. (142f.)
>
> Unter diesen Bedingungen ist Standardsprache keine homogene Varietät und auch kein Bündel von Varietäten, sondern der durch zentrale und prototypische Kennzeichen ausgezeichnete Raum zwischen sprachlichen Leitplanken, deren genaue Setzung dem gesellschaftlichen Interessendiskurs (natürlich nicht völlig frei) anheim geblieben ist. Man kann das auch anders ausdrücken: die Standardvarietät ist nur dann noch eine Varietät, wenn an den verschiedenen Stellen eine Vielfalt von Varietät eingebaut ist. (143)
>
> [...] Der durch die Textsorten und die Kommunikationssituation gesetzte Rahmen ruft durchaus verschiedene sprachliche Eigenheiten hervor, die mit diesem standardsprachlichen Kern als kompatibel angesehen werden. (ebd.)

Elspaß (2005a):

> Historisch gesehen liegt einer Standardsprache ein hohes Maß an Variation zugrunde: Sie ist aus verschiedenen regional, sozial und funktional begrenzten gesprochenen und geschriebenen Sprachen entstanden [...]. Standardsprache"

ist die heute in der Sprachwissenschaft übliche Bezeichnung für die Sprachform, die in einer Sprachbevölkerung überregional und über alle gesellschaftlichen Schichten hinweg akzeptiert ist. (294)

Als wichtiges Kennzeichen einer voll ausgebildeten Standardsprache wird häufig die Kodifizierung von Wortschatz, Grammatik und Rechtschreibung dieser Sprache gesehen. (297)

Elspaß (2005b):[2]

Ich bediene […] mich […] einer 'internationaleren' Definition von Standard:

„Any vernacular (language or dialect) may be 'standardized' by being given uniform and consistent norm of writing that is widely accepted by its speakers. It may then be referred to as a 'standard' language." (Haugen 1994: 4340)

Die Unterschiede zwischen diesen beiden Definitionen sind augenfällig:

- Bei Bußmann ist von einer Normierung auf gesprochener und geschriebener Ebene die Rede, bei Einar Haugen nur von einer einheitlichen *Schreib*norm. […].
- Außerdem fällt die sozialschichtenspezifische Einschränkung in der deutschen Definition auf, durch die offenbar rein deskriptive nicht mehr recht von wertenden Kriterien zu trennen sind […]. Die Rede von „speakers" bei Haugen ist dagegen *schichtneutral*.
- Ein notwendiges Kriterium für Standardsprache bei Haugen, das im „Lexikon der Sprachwissenschaft" fehlt, ist das der breiten *Akzeptanz* der Schreibnorm in der Sprachgemeinschaft. (64)

Glinz (1980):[3]

Ich möchte demgegenüber die gesprochene Sprache ausdrücklich einbeziehen […]. Es scheint mir gerade der Vorteil der Bezeichnung „Standardsprache" zu sein, daß sie den schriftlichen und den mündlichen Gebrauch zusammenzufassen gestattet […]. Dabei schlage ich eine Definition ex negativo vor: unter deutscher Standardsprache der Gegenwart verstehe ich die heute gehörte und gelesene, gesprochene und geschriebene deutsche Sprache, soweit sie als allgemein gebraucht, als nicht-mundartlich und als nicht-schichtenspezifisch betrachtet wird. (610)

Jäger (1980):

[Der zu benennende Gegenstand lässt sich umreißen als] die überregional gebräuchliche Sprache des größten Teils der Gebildeten einer Sprachgemeinschaft, insbesondere aber deren geschriebene Sprache. (376)

[2] Elspaß bezieht sich u.a. auf eine Definition Bußmanns, die weiter oben in der unveränderten Version von 2008 zitiert wird.

[3] Glinz bezieht sich auf die weiter unten zitierte Definition Jägers (1980).

> Nur vereinzelt wird auch gesprochene Sprache der Gebildeten, wenn sie sich durch eine große Nähe zur geschriebenen Sprache (Literatur-, Zeitungssprache usw.) und das Fehlen von Regionalismen auszeichnet, als (gesprochene) Standardsprache bezeichnet. (ebd.)
>
> Da eine genaue Abgrenzung von 'gesprochener Standardsprache' und 'überregionaler Umgangssprache' weder möglich noch nützlich ist, empfiehlt es sich, den Terminus 'Standardsprache' der geschriebenen Sprache vorzubehalten. (377)
>
> Als Standardsprache wird die Sprache bezeichnet, die im Sprachverkehr der oberen und mittleren sozialen Schichten verwendet wird. Aus dem sprachlichen Verhalten dieser Schichten lässt sich die Sprachnorm dieser Schichten ableiten als derjenige Teil ihrer sprachlichen Gesamtkompetenz, der im allgemeinen die Grundlage für ihren Sprachverkehr bildet. (ebd.)

Lüdtke/Mattheier (2005):

> *Standardsprache* ist diejenige Leitvarietät im Sinne eines „idioma cardinale" (Dante), die eine institutionalisierte Verbindlichkeit in Normfragen aufweist. Die teilweise mit ihr konkurrierenden Begriffe *Hochsprache*, *Literatursprache*, *Schriftsprache* [...] oder auch *Nationalsprache* greifen zu kurz. Die Standardisierung erarbeitet über eine Teilkodifizierung eine Varietät heraus, die normativen Charakter beansprucht. Die Standardsprache resultiert aus ihrer Geschichte. (15)

## Quellen

Althaus, Hans Peter/Henne, Helmut/Wiegand, Herbert Ernst (Hg.): Lexikon der Germanistischen Linguistik. 2. Aufl. Tübingen: Niemeyer.

Ammon, Ulrich (2005): Standard und Variation: Norm, Autorität, Legitimation. In: Eichinger/Kallmeyer (Hg.), 28-40.

Barbour, Stephen/Stevenson, Patrick (1998): Variation im Deutschen. Soziolinguistische Perspektiven. Berlin/New York: de Gruyter.

Barbour, Stephen (2005): Standardvariation im Deutschen und im Englischen: Auswirkungen auf die Kommunikation zwischen Sprechern beider Sprachen. In: Eichinger/Kallmeyer (Hg.), 324-333.

Bußmann, Hadumod (Hg.) (2008): Lexikon der Sprachwissenschaft. 4., durchges. u. bibliogr. erg. Aufl. Unter Mitarbeit von Hartmut Lauffer. Stuttgart: Kröner.

Duden (1999): Das Große Wörterbuch der Deutschen Sprache in zehn Bänden. Bd. 8. Hrsg. vom Wissenschaftlichen Rat der Dudenredaktion. 3. Aufl. Mannheim: Dudenverlag.

Durell, Martin (1999): Standardsprache in England und Deutschland. In: Zeitschrift für germanistische Linguistik 27, 285-308.

Eichinger, Ludwig M. (2005a): Standardnorm, Sprachkultur und die Veränderung der normativen Erwartungen. In: Eichinger/Kallmeyer (Hg.), 363-381.

Eichinger, Ludwig M. (2005b): Norm und regionale Variation. Zur realen Existenz nationaler Varietäten. In: Lenz/Mattheier (Hg.), 141-162.

Eichinger, Ludwig M./Kallmeyer, Werner (Hg.) (2005): Standardvariation. Wie viel Variation verträgt die deutsche Sprache? Jahrbuch 2004 des Instituts für Deutsche Sprache. Berlin/New York: de Gruyter.

Elspaß, Stephan (2005a): Zum sprachpolitischen Umgang mit regionaler Variation der Standardsprache in der pluralistischen Sprachgesellschaft. In: Kilian, Jörg (Hg.): Sprache und Politik. Deutsch im demokratischen Staat (= Thema Deutsch 6). Mannheim/Leipzig/Wien/Zürich: Dudenverlag, 294-313.

Elspaß, Stephan (2005b): Standardisierung des Deutschen. Ansichten aus der neueren Sprachgeschichte 'von unten'. In: Eichinger/Kallmeyer (Hg.), 63-99.

Glinz, Hans (1980): Deutsche Standardsprache der Gegenwart. In: Althaus/Henne/Wiegand (Hg.), 609-619.

Haugen, Einar (1994): Standardization. In: Asher, Ron E. (Hg.): The Encyclopedia of Language and Linguistics. 12. Bde. Oxford u.a. Bd. VIII, 4340-4342.

Jäger, Siegfried (1980): Standardsprache. In: Althaus/Henne/Wiegand (Hg.), 375-379.

Lenz, Alexandra N./Mattheier, Klaus (Hg.) (2005): Varietäten – Theorie und Empirie. Frankfurt a.M.: Lang.

Löffler, Heinrich (2005): Wieviel Variation verträgt die deutsche Standardsprache? Begriffsklärung: Standard und Gegenbegriffe. In: Eichinger/Kallmeyer (Hg.), 7-27.

Lüdtke, Jens/Mattheier, Klaus J. (2005): Variation – Varietäten – Standardsprachen. Wege für die Forschung. In: Lenz/Mattheier (Hg.), 13-38.

**Raum für**

**Tauschanzeige**

***amades***